12-2003

Assessment center construct validity : establishing expectations based on the dimension activation theory

Michelle A. Bush

To the Graduate Council:

I am submitting herewith a dissertation written by Michelle A. Bush entitled "Assessment center construct validity : establishing expectations based on the dimension activation theory." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial and Organizational Psychology.

Robert T. Ladd, Major Professor

We have read this dissertation and recommend its acceptance:

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Michelle A. Bush entitled "Assessment Center Construct Validity: Establishing Expectations Based on the Dimension Activation Theory." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial/Organizational Psychology.

Dr. Robert T. Ladd, Major Professor

We have read this dissertation and recommend its acceptance:

Dr. Lawrence James, Committee Member

Dr. David J. Woehr, Committee Member

Dr. David Schumann, Committee Member

Acceptance for Council:

Vice Provost and Dean of Graduate Studies

ASSESSMENT CENTER CONSTRUCT VALIDITY: ESTABLISHING EXPECTATIONS BASED ON THE DIMENSION ACTIVATON THEORY

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Michelle A. Bush

December 2003

DEDICATION

This dissertation is dedicated to my husband, Delond, who provided cherished love, encouragement, and laughter at all the right moments. This dissertation is also dedicated to my parents, Alan and Dee Laird, whose ceaseless faith in me has inspired me to always reach a little higher. And last, but not least, I dedicate this to my Lord and Savior, Jesus, without whom I could not meaningfully achieve.

ACKNOWLDEGMENTS

There are several people who provided me with invaluable guidance and support during my dissertation work. To begin, I would like to thank my committee chair and advisor, Dr. Tom Ladd. He has served as my compass throughout this journey. His dedication and uncompromising standard of expectation are worthy of my sincerest respect and appreciation. Thank you, Tom, for the time and work you committed to my progress, and for your invaluable advice.

I am likewise grateful for the contributions of Dr. Larry James, Dr. David Woehr, and Dr. David Schumann, my dissertation committee. Larry's theoretical insights significantly enhanced the cohesion of my research and his enthusiastic support was greatly appreciated. Dave W. continually encouraged me to develop expectations and view results with unique perspectives, and he provided necessary technical advice for which I am appreciative. Moreover, David S. contributed important recommendations for the framing of my ideas as well as theoretical progression in my methodology. I am blessed to have worked with this distinguished group of professionals.

There are several other amazing people who supported me these last five years with their encouragement, kindness, and helpful advice. Maria, Missy, Jillian, Katie, and Betsy – you are precious friends and I treasure you all. There is so much for which I would like to thank you. Kate – you have become a great mentor to me. I have greatly enjoyed working with you and learning from you. And last, but not least, thank you to Carolyn, Elizabeth, June, Jackie, and Glenda, five incredible women whose kind and giving spirits were a true blessing.

ABSTRACT

The purpose of this study was to examine the construct validity of assessment center dimension ratings within the confines of an extended trait activation theory. Specifically, previous findings of high within exercise rating correlations have led researchers to conclude that ratings are affected by halo. Conversely, the extended trait activation theory suggests that high correlations are a function of the different levels of activation potential for various dimensions rated in a given exercise. For dimensions having stronger activation potential, it is expected that high levels of between subject rating variance will evidence discriminant validity. However, it is expected that dimensions with lower levels of activation potential will show lower levels of rating variance. This central tendency is expected to be the source of high within exercise rating correlations.

Performance based dimension ratings for four distinct exercises were gathered from 97 individuals participating in developmental and selection assessment centers with trained assessors serving as raters. Exploratory factor analyses were conducted for each exercise to determine the necessity of one factor (supporting a halo theory) versus more than one factor (supporting the extended trait activation theory). Moreover, dominance analysis revealed the importance of each dimension for predicting overall exercise performance for each exercise.

For these same exercises, 11 subject matter experts familiar with the exercises and dimensions provided ratings of the relative activation potential of each dimension for

each exercise separately. It was expected that the relative dimension activation scale would correlate with the dimension dominance as revealed in the dominance analysis of actual assessment center ratings. Furthermore, it was expected that the mean variance of activated dimensions would be significantly higher than the mean variance of non-activated dimensions in each exercise. These same analyses were used in comparing variance results with a scale of exercise primacy provided by the original exercise creator, as well. This three point scale was expected to correlate with dimension activation ratings and similar results were anticipated.

PREFACE

The purpose of this study is to examine the suitability of contemporary expectations regarding the construct-related validity of assessment center ratings as typically sought within a multi-trait multi-method framework. In particular, expectations for the demonstration of very low within exercise dimension correlations are challenged based on the characteristics of exercise design. When rating variance and design characteristics are examined simultaneously, it is expected that the dimensions that are more likely to be activated within an exercise will show higher levels of rating variance. Those dimensions less likely to be activated will display low levels of variance and thus, prove the causal agents of significant within exercise dimension rating correlations. Findings will ultimately support the need for revised criteria for construct-related validity evidence within the assessment center domain.

TABLE OF CONTENTS

LIST OF TABLES

TABLES PAGE

CHAPTER I

INTRODUCTION

Assessment centers are an important tool for distinguishing among personnel in today's workplace. In fact, due to their increasing popularity, thousands of individuals are assessed using these procedures each year in a wide variety of organizations (Thornton & Byham, 1982). The popularity of their use may be due in part to the unique design and breadth of the procedure. An assessment center is, by design, a method that reduces rater bias or error (Zedeck, 1986). It is a behaviorally oriented procedure in which multiple assessment techniques are used in the evaluation of a candidate's performance. Judgments of performance are combined and/or integrated across situations allowing raters the opportunity to develop a more complete picture of the assessee's capabilities. Within each exercise that is rated, dimensions are used as a means by which the specific behavioral information is integrated into macro categorizations (Zedeck, 1986). Typically, two assessors are used for every one assessee, allowing for multiple perspectives on an individual's job-related performance. Altogether, the comprehensive design of this assessment procedure is an attractive alternative to reliance on traditional assessment procedures as it allows for the observation and examination of critical dimension-based performance across a variety of situations.

However, despite the obvious methodological improvements over more traditional techniques, assessment centers are not without their own set of inconsistencies. One of the main criticisms voiced against the use of assessment centers

is the lack of evidence for construct-related validity. To elaborate, one of the basic assumptions inherent in the assessment center method is that stable behavioral patterns exist and can be evaluated based on dimensional performance across various exercises (Sackett, 1982). However, time and time again, it has been shown that dimension evaluations within exercises are often more highly correlated than dimension evaluations across exercises (Bycio, Alvares, & Hahn, 1987; Zedeck, 1986; Schneider & Schmitt, 1992). In fact, multiple studies have demonstrated that ratings for a given dimension in one exercise may have little or no correlation with ratings for the same dimension in other exercises (Bycio, et. al., 1987; Sackett & Dreher, 1982). Likewise, exercise factors are often found instead of dimension factors in factor analysis of assessment center ratings (Bycio et. al., 1987; Highhouse & Harris, 1993).

That said, the conceptualization of assessment center construct-related validity upon which these results are based fails to consider the full range of assessee behavior across situations. Particularly, it may be the case that considering the rightful influence of situations on dimensional performance is central to understanding the nature of performance and therefore, determining the validity of dimension ratings. As such, the absence of influential exercise effects may in fact be detrimental to the assessment center rating process. In other words, it is possible that the traditional conceptualization of construct-related validity of assessment center ratings as being supported by high within-dimension rating correlations and low within-exercise rating correlations (e.g., MTMM analyses) is inappropriate.

To explicate, current conceptualizations of construct-related validity of assessment center ratings appear to suggest that the presence of exercise effects is

undesirable. However, it has been suggested that the presence of exercise effects does not necessarily run contrary to the basic tenets of assessment center goals. To be more specific, by design, the assessment center procedure makes use of unique situations in sampling dimensional performance. While it is expected that dimension performance will remain somewhat consistent across situations, it is also expected that different situations will bring about different aspects of performance. In fact, the design of assessment centers is based on this tenet as developers generally create exercises that are unique in nature in order to highlight different aspects of behavior. Were situational differences not expected, assessment center exercises could be composed of highly similar situations and exercise effects would be non-existent. As it is, discrepancies in dimension performance across exercise likely provide important information as to true dimension scores for situationally specific behaviors.

To corroborate, one recent review of studies in which variables were manipulated to assess their impact on construct-related validity of assessment center ratings revealed that findings were typically mixed when considering factors influencing the observation, evaluation, and integration procedures. However, construct-related validity was moderated when considering exercise factors, dimension factors, and assessor factors (Lievens, 1998). Therefore, the focus of assessment center construct related validity studies should be aimed at revealing a better understanding the source of exercise influence rather than attempting to eliminate these effects. This will allow a stronger assessment of construct-related validity as well as aid exercise developers in designing exercises that best capture dimensional performance.

To date, there have been some modestly successful attempts in considering the characteristics of situations as a driving force in construct-related validity findings. For instance, using Bem & Funder's (1978) template matching technique, greater cross-situational consistency in ratings has been found across similar situations (Highhouse & Harris, 1993). Likewise, situations judged as similar in form (group discussion versus role play simulation) have demonstrated higher convergence in ratings (Schneider & Schmitt, 1992).

Several theories have been proffered to describe these results. In particular, a prevalent finding in the assessment center literature has been that only a few performance factors or dimensions are necessary in order to explain the majority of variance in assessment center ratings (Kleinmann, Kuptsch, & Koller, 1996; Schmitt; 1977). In addition, it is clear that there is substantial variability in the manifestation and observation of behaviors relevant to different dimensions across exercises (Reilly, Henry, & Smither, 1990; Bycio, Alvarez, & Hahn, 1987). In fact, different behaviors are typically necessary for successful performance as evidenced by rater judgments (Kuptsch, Kleinmann, & Kohler, 1998). Thus, it seems entirely possible that some dimensions are more dominant in some exercises than in others. Fittingly, it would not be surprising to find that within a given exercise, the behavioral manifestations of certain dimensions would dominate others.

Parallel to this theory, Tett and colleagues (2000; 1999; 1998) have proposed the trait activation theory, based in the personality domain, as holding some explanatory power in the diagnosing problems with construct related validity findings. In particular, they contend that situations differ in their trait activation potential, or trait-relevant

situational cues. Therefore, some traits may be more likely expressed in light of certain situational cues than others. The similarity of situations in terms of the traits that they activate should thus influence the level of cross-situational consistency observed by assessors.

It is the purpose of this study to compare the dominant/salient dimension theory and a modified version of the trait activation theory, the Dimension Activation Theory. In particular, it is expected that some dimensions are more likely to be activated than others within a given exercise, and this propensity can be established with subject matter expert ratings of dimension activation potential. In addition, the variable activation of dimensions within and across exercises may impact the construct-related validity observed in dimensional performance. Specifically, discriminant validity may only be observed in exercises in which a given dimension is activated or has a high level of activation potential. In the same exercise, a non-activated dimension (or one low in activation potential) may tend to exhibit less variant ratings, as the situation is less likely to evoke ample behaviors to warrant more extreme ratings. The central tendency of less dominant ratings could conceivably be the source of the exercise effects found in construct-related validity studies of assessment center ratings.

Demonstration of this effect would ultimately alter the statistical conceptualization of construct-related validation studies of assessment center ratings. In particular, rather than treating exercise effects as undesirable, new conceptualizations of construct-related validity for assessment center ratings would necessarily give account for the unique contributions of exercise factors to the determination of dimension ratings.

Above all, it would relax expectations for very low mono-method, hetero-trait correlations as a result of the decreased variance expected for non-activated dimensions.

Within the current study, dimension ratings in a series of managerial selection and developmental assessment centers are examined with regard expected variance and factor structure based on the dimension activation theory. Specifically, results of exploratory factor analyses will first be examined with the ultimate goal of demonstrating the presence of more than one factor in each exercise flagged with multiple activated dimensions. Following this, dominance analysis is utilized to find the relative influence of dimensions in each exercise when considering overall exercise ratings, and subject matter experts are queried to determine the activation potential and primacy of dimensions within exercises. Similarity in expert ratings of activation potential, dimension design primacy, and established dominance of dimensions would prove to be an initial step in validating these theories with regard to assessment center dimensional ratings. Lastly, the variance of non-activated versus activated dimensions is compared within each exercise to assess the effects of activation on the discriminant validity of dimensions.

CHAPTER II

REVIEW OF THE LITERATURE

Assessment Centers: An Overview

The Assessment Center methodology has been around for more than half a century (Haaland & Christiansen, 2002). Typically defined as "a process employing multiple techniques and multiple assessors to produce judgments regarding the extent to which a participant displays selected competencies" (International Task Force, 2000), this process was originally conceptualized for use in the area of employee selection in a study conducted by the Office of Strategic Services aimed at improving results from the selection of intelligence officers to serve during World War II. Nevertheless, it's comprehensive design has been translated to assist decision makers in early identification of managerial talent, developmental planning, identification of training needs, promotional choice, and determination of management succession (Spychalski, Quinones, Gaugler, & Pohley, 1997; Gaugler, Rosenthal, Thornton, & Bentson, 1987). In fact, the assessment center method has so proliferated contemporary business practice that its use has spanned across a diverse assortment of work domains. This impressive list incorporates industrial settings, educational institutions, military branches, government, law enforcement, and many other specialized organizational settings (International Task Force, 2000; Task Force, 1989). Moreover, although assessment centers are typically reserved for the evaluation of managers, their contribution has extended to the assessment of engineers, college students, salespersons, and blue-collar workers, as well as other non-management professionals (Gaugler, et al., 1987).

The instinctual lure of this method for such diverse purposes can be found in the logic and breadth of the design components essential to the development of successful assessment center tools. An assessment center is, by design, a method that reduces rater bias or error (Zedeck, 1986). In particular, assessment centers are a way in which standardized evaluation of candidates can be conducted based on multiple behavioral inputs (International Task Force, 2000; Task Force, 1989). During the process, judgments are made concerning candidate behaviors by manifold trained observers. Such behaviors are observed as the candidate participates in specifically developed simulated situations. In other words, judgments are made regarding predetermined competencies on the basis of behavioral manifestations in diverse work-related situations. These contrived situations, otherwise referred to as exercises, are in essence samples of typical work performed in the relevant assessment position, and they frequently have quite unique behavioral requirements, enabling a range of relevant behaviors to be demonstrated and observed (Zedeck, 1986; Sackett & Dreher, 1982). Exercises are developed in response to information obtained through thorough job analyses. Specific tasks and necessary performance abilities are identified and defined, and subsequently, exercises based on these abilities are created.

As mentioned, assessment centers are unique in that the measurement variable for the raters is not performance on the task itself; rather, behavioral indicators are condensed into macro categorizations identified as dimensions (Zedeck, 1986). Dimensions represent "a set of tasks and behaviors that are similar in features, or the performance of which requires the same or equivalent ability" (Zedeck, 1986, p.280). The difference between tasks and dimensions is one of outcome versus behavior. Whereas a task

signifies what is accomplished, a dimension highlights behaviors enacted in the pursuit of task accomplishment (Thornton & Byham, 1982). As such, the behavioral foundation of dimensions also distinguishes them from traits, which represent underlying personality constructs. Accordingly, dimensions are analogous to the predetermined abilities isolated in job analyses. As behaviors are gathered as examples of dimensional performance, these dimension indicators are then used to form evaluations. Such distinctions have become increasingly important to the understanding of current debate in the assessment center literature.

An additional advantage of the assessment center methodology is a reliance on multiple observers or raters. The use of multiple observers is a design component aimed at reducing cognitive biases that often occur when only one rater is utilized. Moreover, expert raters are provided multiple opportunities to view relevant performance cues across the controlled stimulus settings (Jones, 1992). Beyond this, as an integral element in the assessment center process, individual rater judgments are combined either statistically, or more often in the form of a consensus discussion, with the intention of increasing behavioral representation and rating accuracy (Zedeck, 1986). This process presumably allows raters to develop a more complete picture of the assessee's capabilities. It is then the final dimensional ratings, or the overall assessment ratings that are based on dimensional ratings, that are used to make judgments as to a candidates' suitability for a job and/or developmental needs.

However, the assessment center methodology has gained favor for reasons beyond its demonstration of face validity (Howard, 1997). Quite consistently, assessment center results have established strong levels of prediction in relation to multiple success

variables (Gaugler, et al., 1987; Schmitt, Gooding, Noe, & Kirsch, 1984; Turnage & Muchinsky, 1984; Klimoski & Strickland, 1981). The most notable of such results dates back to the first industrial application of the assessment center by the AT&T Management Progress study (Bray & Grant, 1966; Bray, Campbell, & Grant, 1977). During the course of this study, 422 candidates were assessed on 25 dimensions that were predicted to be related to managerial performance and progression. In order to avoid confounding effects based on information usage, the assessment center results remained inaccessible to decision makers and were used solely for research purposes. It was subsequently found that predictions regarding candidate attainment of management levels based on assessment center results predicted management level achieved five to seven years later with correlations ranging from .44 for college graduates and .71 for non-college graduates. Since the release of this study's results and implementation of assessment centers across additional domains, criterion-related validity evidence has continued to spark confidence. Most notably, Gaugler et al. (1987) conducted a meta-analysis on the validity of assessment center ratings which included results from 50 studies, both published and unpublished, containing 107 validity coefficients. To summarize, they found the average criterion-related validity to be as high as .40 for some outcomes, with an average corrected mean and variance of .37 and .017, respectively. While assessment centers results have been linked most strongly to career potential and advancement criteria in some comparative studies (Turnage & Muchinsky, 1984; Klimoski & Strickland, 1977), others have revealed significant links to performance. As an example, Thornton & Byham (1982) reviewed numerous predictive validity studies analyzing the relationship between overall assessment ratings and performance or

progress criteria. They found that the overall assessment rating was highly correlated with several performance indicators over time including salary, performance ratings, and ratings of potential. More recently, Russell & Domm (1995) conducted an assessment center created to select store mangers for a durable good manufacturing company. District managers unfamiliar with the candidates assessed 140 current store mangers. These authors found considerable links between overall assessment ratings and overall responsibility ratings gleaned from supervisor performance appraisals ($r = .28$), and overall assessment ratings and store profit ($r = .32$). These results were further translated to fiscal projections which revealed that candidates scoring in the top two of four rating points generated $3000 more quarterly profit in their stores than those scoring among the bottom two points. Such convincing results have undeniably influenced the popularity of this assessment method.

Nevertheless, the appeal of the assessment center methodology stretches beyond its attractive design and predictive power. Incidentally, assessment center ratings have been shown to be free from the effect of adverse impact, an unpleasant factor plaguing other successful predictors such as cognitive ability measures (Howard & Bray, 1988, as cited in Howard (1997); Thornton & Byham, 1982). Moreover, when used for developmental purposes alone or in conjunction with selection, assessment centers have the unique advantage of isolating developmental priorities. Particularly, participants are provided detailed feedback regarding their strengths and weaknesses on relevant performance dimensions (Lievens, 2002; Howard, 1997). This feedback can then be applied to subsequent training and career planning initiatives.

The Validity Debate

Despite the obvious methodological improvements over more traditional techniques, the assessment center methodology has not escaped concerted criticism. In general, the foundation of this assessment method is that people are consistent in the pattern of their behaviors, and that behaviors elicited in assessment center exercises can be meaningfully categorized into the relevant performance dimensions (Sackett & Dreher, 1982). In view of that, assessment center exercises represent opportunities for the demonstration and observation of behavior relevant to the predetermined dimensions. In essence, assessment center exercises are designed according to those dimensions deemed important via job analyses, and dimensional performance, rather than exercise performance, becomes the construct of interest. Typically, assessors observe and record candidate behaviors across numerous exercises, classify those behaviors into relevant dimensions, and evaluate each candidate on each dimension (Gaugler & Thornton, 1989). Following this, consensus discussions are used to integrate assessor evaluations to form conclusions regarding candidate potential or developmental needs in terms of the dimensions. Again, the focus of the conclusions and evaluative feedback is on performance within the dimensions, not within exercises or tasks.

That said, the results of numerous construct related validity studies involving assessment center ratings have led many researchers to question the traditional belief that assessment center ratings represent dimensional performance and postulate that instead, they may be more reflective of general performance within exercises or performance tasks (Sackett & Tuzinski, 2001; Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000; Lowry, 1997; Schneider & Schmitt, 1992; Bycio, Alvares, & Hahn, 1987;

Klimoski & Brickner, 1987; Robertson, Gratton, & Sharpley, 1987; Neidig & Neidig, 1984; Sackett & Dreher, 1982; Neidig, Martin, & Yates, 1979; Sackett & Hakel, 1979; Archambeau, 1978; Klimoski & Strickland, 1977). In other words, it has often been argued that although assessment center ratings appear to demonstrate both content validity and criterion-related validity, evidence suggests that ratings do not demonstrate construct-related validity. Principally, concern lies with findings that indicate that assessment center ratings do not represent separate constructs.

These conclusions have most typically followed from studies employing Campbell and Fiske's (1959) multi-trait, multi-method approach in evaluating dimension and exercise effects. Using this method, average within dimension rating correlations are compared with average within exercise rating correlations to assess the relative influence of dimension and exercise factors. The majority of studies utilizing this method have found higher within exercise rating correlations than within dimension rating correlations, with the average within dimension rating correlations generally ranging between .16 and .36, and the average within exercise rating correlations generally ranging between .41 and .75 (Haaland & Christiansen, 2002; Sackett & Tuzinski, 2001; Fleenor, 1996; Joyce, Thayer, & Pond, 1994; Harris, Becker, & Smith, 1993; Schneider & Schmitt, 1992; Reilly, Harris, & Smither, 1990; Sackett & Harris, 1988; Bycio, Alvarez, & Hahn, 1987; Klimoski & Brickner, 1987; Russell, 1987; Sackett & Dreher, 1982). This pattern has been interpreted as evidence for a lack of convergent and discriminant validity among dimension ratings, though the majority of concern seems to lie with purported inability to distinguish among dimensions.

Nevertheless, this approach to understanding the construct-related validity of assessment center ratings has similarly been questioned. Most importantly, the multi-trait, multi-method framework is limited in that it does not specify appropriate criterion for evaluating the model most suitable for estimating effect sizes of dimensions and exercises (Schmitt & Stults, 1986; Widaman, 1985). In particular, these studies face scrutiny as the statistics are based on correlations among observed variables, which ultimately contain measurement error (Kleinmann & Koller, 1997). As such, the coefficients are a function of the reliability of the variables. In addition, rater effects and exercise effects are often confounded in these studies as they are treated as one and the same (Jones, 1992). Therefore, researchers have more recently turned to confirmatory factor analysis as an alternative method for assessing construct-related validity. This method has significant advantages over the multi-trait multi-method framework in that it considers underlying constructs rather than relying only on observed variables, models can be compared for appropriateness, and dimension and exercise variance can be examined (Lievens & Conway, 2001).

That said, results of studies employing the confirmatory factor analysis methodology have likewise been less than optimal. Exercise variance appears to yet again dominate dimension variance of ratings in these models. For instance, Bycio, Alvarez, & Hahn (1987) used confirmatory factor analyses to examine assessment ratings for 1170 manufacturing supervisors and supervisor candidates over a period of five years. Within this assessment center, eight dimensions or abilities were measured in five situational exercises. They found that exercise variance dominated the confirmatory factor analysis with exercises contributing more variance than ability and error combined.

Moreover, in multiple studies comparing alternative models, those including multiple exercise constructs, but only one general performance factor (rather than separate dimension factors), have at times provided the best fit with assessment center rating data (Lance, et al., 2000; Schneider & Schmitt, 1992). This would suggest that raters are unable to distinguish among various assessment center dimensions when making ratings.

Nevertheless, confirmatory factor analysis studies have shed some optimistic light on the influence of dimensions in the rating process, as well. That is, despite the high levels of consistency in ratings found within exercises, there remains evidence that the dismissal of dimensions from the rating process would be erroneous (Sackett & Harris, 1988). To be more specific, multiple studies have found that the inclusion of dimension factors is vital for establishing good model fit. As an example, in a study involving the assessment of 190 candidates for a police promotional examination, researchers compared competing models with regard to dimension ratings. Findings indicated that a model including nine oblique trait (dimension) factors and four oblique method (exercise) factors provided better fit with the data (CFI = .96) than models including only exercise factors (CFI = .81 to .89) or including exercise factors along with one general performance factor (CFI = .90 to .92) (Donahue, Truxillo, Cornwell, & Gerrity, 1997). Likewise, Kudisch, Ladd, & Dobbins (1997) used confirmatory factor analysis to examine the appropriateness of four competing models: a) a model including seven rated dimensions and four exercise factors as indicated by the inherent design of the assessment center, b) a model including one general dimension factor and four exercise factors as proposed by previously reviewed authors (Schneider & Schmitt, 1992; Bycio et al., 1987), c) a model including solely four exercise factors, and d) a model including

solely seven dimension factors. Although a slight modification was necessary to obtain model convergence (i.e., the analysis and judgment dimensions were combined), the model including separate dimension and exercise factors provided the strongest fit with the data (CFI = .941). In addition, low correlations between the latent dimension factors (e.g., ranging from -.23 to .65) suggested that assessors were able to discriminate between abilities; though, the exercise variance still exceeded that attributable to dimensions (mean variance = .41 and .25, respectively).

Altogether, though, most recent evidence regarding the presence of dimension effects in assessment center ratings has seemingly not been strong enough to conclusively dispel doubts regarding the appropriate use of and reliance upon assessment center dimensions. In the relevant literature, numerous deductions based on these results have been grim: "The bulk of the reported literature shows little support for the view that assessment center procedures do in fact produce scores that serve as valid representations of separate constructs or that those constructs are used in evaluation decisions in the manner proposed by the assessment center designers" (p.245, Klimoski & Brickner, 1987). Moreover, calls have been placed to go as far as altogether eliminating dimensions from the assessment center framework (Lowry, 1997; Robertson, et al., 1987), and suggestions have been advanced to reconceptualize the assessment center process as a series of miniature work samples designed to elicit job relevant behaviors (Robertson, et al., 1987; Silverman, Dalessio, Woods, & Johnson, 1986; Neidig & Neidig, 1984; Sackett & Dreher, 1982).

On the other side of the debate, some have compellingly argued that the mere presence of content and criterion-related validity and absence of construct-related validity

is paradoxical when considered within a unitarian framework of validity (Arthur, Woehr, & Maldegen, 2000). Within this perspective, content-, criterion-, and construct-related validity are not orthogonal; instead, they are simply different ways to measure or demonstrate construct-related validity of a measure (Binning & Barrett, 1989). Thus, if any two strategies are successfully established, conceptually, demonstration of the third is a requirement.

It has also been suggested that construct-related validity evidence should be questioned in that the characteristics of assessment centers make it difficult to accurately evaluate their construct-related validity (Jones, 1992). Most importantly, despite concerted attempts at standardizing the assessment center process (International Task Force, 2000; Task Force, 1989), assessment centers tend to be widely variant across organizations (Zedeck, 1986; Schmitt, Schneider & Cohen, 1990). To illustrate, in a comprehensive meta-analysis conducted by Gaugler et al. (1987) it was determined that assessment centers differed significantly in the number of assessment devices utilized (range = 1 to 11), days of observation (range = 1 to 3), and assessor type (managers, psychologists, or both). In addition, variant rating procedures have been employed (dimension- versus exercise-based), and integration procedures have not received ample research attention. Furthermore, dimensions differ in their specificity versus generality, and fuzziness versus clarity (Guion, 1987). In summary, generalizing construct-related validity findings across organizations is difficult as differences in findings may be due to situational differences (Schmitt, Schneider, & Cohen, 1990).

In order to examine these assertions, Arthur and colleagues (2000) developed an assessment center emphasizing many of the features and recommendations suggested by

research to improve the convergent and discriminant validity of dimension ratings (e.g., limitation of dimensions to nine; 2 to 1 assessor to candidate ratio; use of psychologists as assessors; frame-of-reference training prior to assessment). The developmental assessment center was conducted with 149 government employees and included four exercises. Results based on generalizability analysis showed that dimension main effects and the person by dimension interaction accounted for significant portions of the variance in ratings (21% and 20%, respectively). These figures greatly exceeded those found in relation to an exercise main effect or person by exercise interaction, which accounted for less than 1% and approximately 5% of the total variance, respectively. Likewise, the confirmatory factor analysis outcomes revealed that dimension parameter estimates exceeded those for exercises in all cases. Specifically, the mean dimension parameter estimate was .77 and mean exercise parameter estimate was .26. Thus, the future outlook for dimensions within exercises may not be as dismal as once projected.

Nevertheless, it cannot be argued that exercise effects do not have a impact on dimension ratings that is stronger than expected based on statistical theory related to the multi-trait, multi-method analysis (Campbell & Fiske, 1959) and confirmatory factor analysis (Marsh & Grayson, 1995). In addition, the improvements in construct-related validity found when altering assessment center methodology does not conclusively address the questions proposed by Klimoski and Strickland (1977) or Klimoski and Brickner (1987) regarding whether or not assessment center ratings actually reflect the dimensions they are purported to measure (Joyce et al., 1994).

This issue has become prevalent in the assessment center literature and has been labeled as having vital implications for the future use and acceptability of the assessment

center methodology for various purposes. Most notably, concern lies with the use of assessor ratings of performance within dimension as tools for developmental feedback in diagnostic assessment centers (Lievens & Conway, 2001;Kudisch, et al., 1997; Fleenor, 1996; Joyce, et. al., 1994; Bycio, et al., 1987; Thornton & Byham, 1982). As dimensions are typically used to provide performance feedback, the information contained in feedback reports and subsequent actions could have detrimental effects if dimensions are not actually valid constructs of performance. Moreover, the construct validity of dimensions has repercussions for the criterion-related validity of assessment center ratings. To expound, when more accurate dimension ratings are utilized when assigning final ratings, prediction accuracy will be increased (Thornton & Byham, 1982). Whereas we may never be able to conclusively estimate the accuracy of assessment center ratings, as there is no true score with which to compare them, establishment of rating reliability and validity may be sufficient (Sulsky & Balzer, 1988).

Relevant Research in the Personality Domain

All in all, the establishment of confidence in the use of dimension specific ratings appears to be critical for the continued viability of assessment centers as we know them. However, the issues that have become central in assessment center debate are by no means new to academic literature. In order to best understand the nature of the debate in the assessment center literature regarding the appropriate application of dimensions, it is necessary to consider parallel considerations that have proliferated personality research for decades. The core of the person-situation debate of personality has similarly centered on the existence of behavioral consistency and thus, the reality of traits. To be more specific, it has often been demonstrated that correlations between objective measures of

the same trait are low, and that greater amounts of variance in behavior have been accounted for by situations and person-situation interactions than by person factors alone (Epstein & O'Brien, 1985). Though at one time the debate actively vacillated between those supporting a situationist perspective (there are no such things as traits) and those supporting a trait perspective, there currently exists a consensus that the interaction between person factors and situational variables is where the action takes place (Bem & Funder, 1978; Bem & Allen, 1974; Magnusson, 1982). More importantly, identification of temporal stability in behavioral patterns has emerged as a key factor in the person-situation arena (Epstein & O'Brien, 1985;Mischel & Peake, 1982). In fact, it has been proposed that behavioral variation related to differing situations provides meaningful information about the personality system (Mishel & Shoda, 1998). Specifically, individual differences in the patterns by which behavior varies across situations represents coherence of personality (Shoda, Mischel, & Wright, 1993). In other words, according to assumptions based on the cognitive affective personality system model, the relationship between situations encountered and resulting cognitions, affects, and behavior is determined by the personality system. Thus, the information gathered from changes in behavior as a function of situation is not a source of error to be eliminated, but instead, a key to understanding person characteristics (Mishel & Shoda, 1998).

As the personality literature moves toward revealing a better understanding of the impact of situational differences on behavioral choice, a shift or addition in focus is likewise necessary within the assessment center construct-related validity research domain. In particular, it is imperative that research efforts aimed at eliminating situational contributions to rating variance be supplemented with an increased focus on

expanding current understanding of the process underlying the these effects (Zedeck, 1986). Chiefly, the lack of systematic research regarding situational impact on dimension-based behavior has left exercise designers at a disadvantage in their attempts to develop simulations that most accurately sample situational aspects of target jobs (Neidig & Neidig, 1984; Schneider & Schmitt, 1992; Bycio, et al., 1987). To explicate, in addition to current concentration on demonstrating the convergent validity of assessment center dimension ratings, more attention should be focused on delineating what happens within relevant situations (assessment center exercises) that is precluding the consistent manifestation of discriminant validity between the dimensions as defined by current statistical methodology.

Construct Related Validity in the Generalizability Framework

Such a focus on construct-related validity is supported when the previous problems are considered in light of generalizability theory. To clarify, generalizability studies work to partition variance estimates into their underlying causal components and interactions among those same components (Cronbach, Gleser, Nanda& Rajartnam, 1972). Principally of interest in assessment center construct related validity studies are situational or exercise variance components, person or subject variance components, dimension variance components, and all subsequent interactions. That said, in a recent generalizability study that considered each of these factors, it was revealed that the person by exercise by dimension interaction accounted for up to 44% of the total rating variance, an amount greater than the variance contributions of all other factors combined (Bush & Ladd, 2002). Therefore, it is of central interest to assessment center researchers to develop an understanding of the nature of this interaction.

Explanatory Theories for Construct Validity Results

To this end, several theories have been advanced as holding explanatory power with regard to this phenomenon. Generally, these theories stem from fairly consistent findings that overall assessment center ratings are typically based on judgments regarding candidate performance on relatively few dimensions (Lance et al., 2000; Fleenor, 1996; Kleinmann, Kuptsch, & Koller, 1996; Russell, 1985; Sackett & Hakel, 1979; Schmitt, 1977). To illustrate, Sackett & Hakel (1979) used multiple regression and factor analysis to examine decision strategies of assessors. They examined judgments for 719 assessment center candidates and found that two factors, leadership and organizing-planning/decision-making, were common to assessors and identified as important by the regression model. Likewise, in attempts to partially replicate and extend this study, Russell (1985) directed assessors to equally weigh four apriori categories of dimensions when making overall assessment ratings. Findings indicated that a single factor dominated initial dimension ratings and the apriori categories did not reflect the most accurate grouping of categories underlying the dimension ratings. Similar to Sackett & Dreher's (1982) findings, only three dimensions accurately predicted overall assessment ratings: leadership, organization and planning, and decision-making.

Predominantly, it appears that researchers have initiated attempts to understand these results in terms of the similarity and/or dissimilarity of *dimensions* used in a typical assessment center; in other words, attention is not focused on the exercise factor. More specifically, these theories have endeavored to dissect the differences among dimension characteristics in order to explain why some dominate others across situations. For instance, one such study investigated the effects of the transparency of dimensions on the

construct validity of ratings (Klienmann, Kuptsch, & Koller, 1996). One hundred nineteen college students participated in an assessment center where dimensions were either communicated to them prior to the assessment or not communicated at all. With the transparency condition (where candidates know what dimensions and behavior was required of them) construct validity was higher than within the non-transparency condition. Particularly, three ability factors and three exercise factors were included in the best fit model via confirmatory factor analysis of the transparency condition, but only one ability factor (oral communication) and three exercise factors were necessary in the non-transparency condition.

Still others have suggested that the answer may be found in the specificity versus breadth of dimensions and behaviors which comprise them (Joyce, Thayer, & Pond, 1994; Guion, 1987). In particular, it has been proposed that the inclusion of higher order constructs or broader attributes may be necessary to demonstrate construct-related validity in ratings. With narrow constructs, fewer behaviors are typically relevant to performance ratings within each dimension, and thus fewer relevant behaviors would be observed.

On the other side of the fence, paralleling more recent studies in the area of personality which have shown that the characteristics of situations have a robust impact on the observation of cross-situational consistency (Mischel & Shoda, 1995; Shoda, Mischel, & Wright, 1993), researchers have attempted to explain the exercise effects in terms of *situational* similarity and/or dissimilarity (Neidig & Neidig, 1984). To expound, Brannick, Michaels, & Baker (1989) examined the convergent and discriminant validity of data collected on parallel in-basket forms using multi-trait, multi-method analysis.

They found weak convergent and discriminant validity results both within and between the two forms. Reliability estimates of ratings between forms ranged between .21 and .43. Thus, even situations that appear to draw on the same abilities or traits may lack formidable amounts of shared exercise variance (Sackett & Dreher, 1982). That said, several other studies have found evidence that cross-situational consistency is higher when situations are more similar (Highhouse & Harris, 1993; Lord, 1982; Sackett & Harris, 1988; Shoda et al, 1993).

In a recent attempt to test the situational specificity hypothesis against one of method bias, Lance, Foster, Gentry, & Thoresen (in press) considered the expected correlation of scores with performance criteria based on the two contingencies. Specifically, using confirmatory factor analysis model that included a relevant job knowledge measure, they found a correlation between the job knowledge performance factor and each of the exercise factors ($r = .16$ to $.24$), as well as a general performance factor ($r = .36$). These results support the assertion that exercise effects typically found in construct related validity studies of assessment ratings may not constitute unwanted performance-irrelevant error.

That said, it is yet unclear how the exercise effects operate and why this may occur (Tett & Guterman, 2000). The majority of recent theorizing regarding these effects has concentrated on examining the dynamics involved with the interaction of dimensions and exercises. The preponderance of the discussion on this issue has resulted in postulations that some dimensions are more dominant than others within exercises. In general, two explanations have generated noteworthy attention: Dominant/Salient Dimension Theory and Trait Activation Theory.

Dominant /Salient Dimension Theory

The dominant dimension/salient dimension hypothesis primarily involves consideration of the relative salience of dimensions within each exercise. This class of theories purports that there are wide variations in the opportunity to manifest behaviors representative of certain dimensions and/or opportunities to observe behaviors of different dimensions within exercises. Somewhat in support of this theory, it has been established that wide variations exist in the opportunity to display dimension-relevant behavior across exercises (Donahue, Truxillo, Cornwell, & Gerrity, 1997), and therefore, it may not be prudent to treat all dimensions as equal. Most telling, one study involving the listing of observed behaviors for various dimensions found that that number of behaviors observed ranged from 4 in one dimension to 32 in another (Reilly, Henry & Smither, 1990). Nevertheless, while some have argued that the cause of sub-optimal construct-validity findings is a result of insufficient opportunity to reveal dimensional behavior in individual exercises, diligent assessment center design should preclude this sort of effect. Appropriate assessment center design does not advance the consistent application of dimension ratings in exercises in which they are not relevant or important. Yet, even with sufficient opportunities to observe behaviors relevant to each dimension assessed in each exercise, it is probable that a given dimension may be more observable in some exercises than in others (Sackett & Dreher, 1982). In fact, it is in harmony with assessment center design to anticipate such differences. Mainly, as per Douglass Bray, originator of the first management assessment center, exercises are expected to represent major domains (Howard, 1997). Therefore, it should not be expected that each and every exercise should equivalently elicit each and every dimension that is rated; rather, it is

logical to suggest that exercises are designed with certain dimensions dominating. For instance, a leaderless group discussion may be designed with the aim of uncovering leadership abilities to a greater extent than planning and organizing skills. Likewise, an in-basket exercise may be created with the intention of uncovering planning and organizing skills to a greater extent than customer orientation skills.

This understanding is particularly important to the discussion of construct-related validity in that within this class of theories. It has been proposed that ratings of salient dimension in a given exercise may influence subsequent ratings of less salient dimensions within the same exercise (Lance, et al., in press; Lance, LaPointe, & Stewart, 1994). In other words, the dominant/salient dimension model proposes that a level of halo exists within dimension ratings, resulting from the influence of a salient dimension on evaluations of behavior in non-salient dimensions, and low levels of convergent and discriminant validity are a result of this halo. Support for this theory would leave assessment center users at a disadvantage. Particularly, the logical consequence would be that only for salient dimensions could meaningful interpretations of behavior be advanced. Ratings and interpretation of behavior based on non-salient dimensions would subsequently be of limited use to decision makers or assessment center participants.

Despite numerous attempts, empirical support of this theory has not been firmly established. Of primary relevance, Lance and colleagues (in press) provided a test of the salient dimension hypothesis as contrasted with a general impression hypothesis with regard to same exercise, different dimension performance. After identifying salient dimensions within three assessment center exercises (as per assessor judgments), confirmatory factor analysis was used to compare models in which: a) the nonsalient

dimensions were regressed on the salient dimension (the salient dimension model), b) both salient and nonsalient dimensions were regressed on an overall general impression rating for the exercise (the general impression model), and c) the salient dimension was regressed on the general impression factor and nonsalient dimensions were regressed on both the salient dimension and the general impression factors (the combined model). Results indicated that the combined model provided the best fit, though the general impression model outperformed the salient dimension model.

Of note, this theory, along with most proffered in the assessment center literature, fails to adequately consider the decades of research highlighting the interaction of person and situation factors in the explanation and interpretation of behavior. Thus, it is not surprising that new theories emerging out of the personality literature are recently receiving concerted attention in the assessment center literature. Although it harbors some similarity with the dominant/salient dimension theory, the principle of trait activation attempts to shed a revealing light on the dilemma.

Trait Activation Theory

Returning to the personality literature, the principle of trait activation in relation to assessment center ratings as proposed by Tett (1998, 1999, as cited in Lievens & Klimoski, 2001) and Tett and Guterman (2000) has been increasingly used to describe the interaction between personality and situations. Specifically, this theory argues for situation trait relevance, or opportunity to express a trait, as a moderator of trait-behavioral relations and cross-situational consistency. Said differently, the expression of personality traits requires trait-relevant situations (Kenrick & Funder, 1988), and only

when two situations share trait-expressive opportunities can cross-situational consistency in ratings be anticipated (Tett, 1998, 1999 as cited in Lievens & Klimoski, 2001).

As exemplified by Tett & Gutterman (2000, p. 398), "personality traits are intraindividually consistent and interindividually distinct propensities to behave in some identifiably way...people high on aggression, for example, do not always behave aggressively; they do so only in certain situations. The question is, which situations". To illustrate, consider typical behavior observed at a religious event. It would be unusual to witness high levels of variability in aggression as religious proceedings offer few cues for the trait expression (Tett & Guterman, 2000).

Researchers investigating this interaction have primarily focused on trait-intention correlations. Specifically, in a study conducted by Tett and Guterman (2000), the researchers targeted five traits to be elicited by two exercises each in five different life domains, totaling fifty exercises. In addition to completing self-report personality measures relevant to the five specific traits, subject were presented with the fifty scenarios and asked to report their intended response in each situation. For three out of five traits, trait-intention correlations were indeed higher for more relevant situations and cross-situational consistency in intention scores was higher for situations similarly high in target trait relevance; though, these results were not uniform across the board (e.g., they ranged from -.02 to .39 for the trait of risk-taking).

The trait activation theory has been identified as especially relevant to the assessment center domain as it potentially provides a link between the personality/situation interaction and performance within exercises that has previously been ignored within the academic literature. However, in the original trait-activation

theory, trait relevance represents a qualitative situational feature that allows for differences in the expectation of trait expression (Tett & Burnett, 2002). When applied to the assessment center methodology, performance within exercise is expected to be a function of the expression of such traits. That said, contrary to the above-mentioned study, assessment center exercises are not designed to be relevant for specific personality traits. Rather, they are designed directly around performance dimensions. Nevertheless, the theory potentially holds significant explanatory power in the assessment center domain. The trait activation theory asserts that it should not be expected that each and every situation provide equal trait-relevant situation cues. Likewise, it should not be expected that each exercise should equivocally cue each and every dimension that is rated. In fact, it may be that as a result of unique dimension-relevant cues within exercises, dimensions engender differing levels of activation potential themselves.

In discussing these differences, it is imperative to point out the importance of maintaining the integrity of performance dimensions as traditionally conceptualized rather than considering performance within exercises as strictly a function of trait expression. Performance dimensions may in fact hold the key to the success of the assessment center methodology. Particularly, any attempt to measure and give feedback regarding performance-related behavior in terms of personality variables during assessment center exercises would require significant levels of inference on the part of assessors and resulting feedback would have limited applicability with regard to anticipating performance in certain situations. On the other hand, the use of behaviors and a classification system of dimensions precludes the necessity of heavy inferences and provides prospective companies and candidates with specific behavioral indicators of

future performance in like situations, as well as a focus for developmental efforts. Finally, as assessment center exercises are designed to elicit behavior relevant to certain performance dimensions, not traits, it is reasonable that our criterion of interest is dimensional performance, not trait expression.

In summary, rather than considering assessment center construct related validity in terms of the trait activation theory, it is suggested that the same principals be applied to assessment center exercises in terms of the activation of dimensions. In other words, the Dimension Activation Theory holds that performance within exercises will be a function of the enactment of certain behaviors relevant to some dimensions more so than others. When a dimension of performance is activated by situational cues (as per the exercise design purpose), behaviors relevant to that dimension are expected to be observed by raters. In the case that a dimension of performance is not highly activated, fewer performance indicators are expected.

In Comparison

Presently, it is important to note a critical difference between the dimension activation theory and the previously reviewed dominant/salient dimension model. In particular, the dominant/salient dimension model suggests that one or two dimension(s) may dominate the rating process as per purported salience, and ratings of the said dimension(s) would effect evaluations in dimensions purported as less salient. This halo in dimension ratings would consequently manifest itself in high within exercise rating correlations, or in other words, a lack of discriminant validity. Changes between exercises in dimension dominance would seemingly be the source of lowered levels of convergent validity, as well.

Conversely, the dimension activation theory does not suppose such an effect on non-dominant or non-activated dimensions. Rather, according to the latter theory, non-activated dimensions would only show a lower amount of rating variance than activated dimensions. Although the result would likewise be lower levels of discriminant validity, this would primarily be among non-activated dimensions. To explicate, if a situation is not highly relevant for planning and organizing skills to be demonstrated, it is unlikely that an individual will take action indicative of very strong planning and organizing skills or very weak planning and organizing skills unless the skill or lack thereof is paramount in that individual's skill set (that is not to say that relevant or useful information cannot be gathered from consideration of the candidate's planning and organizing behavior in that exercise). Thus, it might be expected that low levels of rating variance within exercises may be found across non-activated dimensions, resulting in high levels of intercorrelations. On the other hand, when certain dimensions are activated, clearly differentiated dimensional ratings may emerge between candidates. As a result, performance on activated dimensions should show significant levels of variation, and thus show discrimination when compared with less activated dimensions.

The Present Study

The purpose of this study is examine assessment center dimension ratings within the confines of the dimension activation theory and the dominant/salient dimension theory in order to develop a more comprehensive understanding of the nature of within exercise dimension ratings. In particular, the structure of dimension activation potential within exercises is compared to that expected in light of the dimension activation theory, and comparisons among the expected variance and factor structure of within exercise

dimension ratings are examined as evidence for theory applicability. An empirical distinction between these two theories is integral to understanding the nature of prior construct-related validity studies and has vital implications for the continued use of dimensions in the rating process. Essentially, a strong fit of the data within the expectations of the dimension activation theory would suggest that current problems with discriminant validity of dimension ratings within exercises lie in the interpretation of results based on standard methodology. Particularly, considerable correlations among non-activated dimensions should be expected. On the other hand, a strong fit of the data within the expectations of the dominant/salient dimension theory would suggest that the low levels of discriminant validity within exercises are actually a result of halo.

Research Hypotheses

To begin, the factor structure of the ratings within each exercise will be considered. Specifically, in contrast to the unidimensionsal dominant/salient dimension model, the dimension activation theory implies that if two or more dimensions are activated, there will be multidimensionality in the model. Consequently, it is first necessary to test the dimension activation theory by determining whether one factor is sufficient to describe the data or multiple factors are necessary. Consistent with the dimension activation theory, it is expected that exploratory factor analysis will reveal the presence of more than one performance factor. Therefore,

H1: Whenever an exercise reflects two or more activated dimensions, the within exercise dimension ratings will represent more than one factor.

Next, to establish coherence with prior construct-related validity studies, multi-trait, multi-method analysis will be performed on a sample of dimension ratings obtained

on candidates participating in managerial selection and developmental assessment centers. Analogous with previous findings, it is expected that:

H2: The average mono-method, hetero-trait correlation coefficient will exceed the average hetero-method, mono-trait correlation coefficient.

Following this, subject matter expert ratings of relative dimension activation potential will be collected within each exercise to distinguish between activated and non-activated dimensions. Two sources of subject matter experts will be queried: the assessment center developer and assessors. Following the method used in exercise creation, the assessment center exercise developer will rate all dimensions for each exercise in terms of their primacy in development. This will be called the design indicator scale.

Likewise, assessors will give direct estimations of the relative activation potential of each dimension in each exercise. As assessors will have received training with regard to dimension definition/behavioral indicators and exercise specifics and will have considerable experience with the rating of the exercises, it is believed that their judgments regarding relative dimension activation potential will be reliable. It is also expected that subject matter expert (assessor) ratings of relative dimension activation potential will correlate with indicators of the contrived dominance of dimensions included as the exercises were being designed (the design indicator scale). More specifically, as exercises were designed, the creator had specific dimensions in mind as the primary behaviors to be elicited in the situation. Other dimensions were included as having a secondary or tertiary role. It is expected that those dimensions the exercise was

designed to elicit will correspond to those dimensions that subject matter experts indicate are most likely to be activated.

H3a: There will be agreement among subject matter expert (assessor) ratings of relative dimension activation potential within each exercise.

H3b: There will be consistency between the exercise designer's ratings of dimension primacy and the subject matter expert ratings of dimension activation potential.

In an attempt to establish the plausibility of the dimension activation theory with regard to assessment center dimension ratings, dimension ratings within exercises will be examined to determine their relative dominance within each exercise. These findings will then be compared to subject matter expert ratings, as well as the design indicator scale. Agreement will be indicative of theory plausibility.

H4a: There will be consistency between subject matter expert ratings of dimension activation potential and the relative importance of dimension ratings in estimating overall exercise performance within each exercise.

H4b: There will be consistency between design indicators of dimension primacy and the relative importance of dimension ratings in estimating overall exercise performance within each exercise.

Finally, while this information will give indication as to assessors' and the designers' beliefs regarding the activation of dimensions within exercises, as well as the tendency of ratings to reflect these beliefs, examination of the variance of the ratings in light of the proposed theory is necessary to provide evidence that the differences between dimensions in activation potential is the source of the construct-related validity results

found in previous studies. Therefore, with the aim of translating these results to the development of a better understanding of exercise effects on the discriminant validity of assessment center dimension ratings, rating variance will be compared among dimensions differing in levels of activation. It is expected that:

H5a: The activated dimensions within each exercise will show higher levels of rating variance than non-activated dimensions within the same exercise.

H5b: The primary activated dimensions as indicated by the design indicator scale will show higher levels of rating variance than the secondary/tertiary dimensions within the same exercise.

CHAPTER III

METHODOLOGY

Overview of the Study

This study represents an initial examination of the dimension activation theory as compared with the dominant dimension theory as a potential explanation of assessment center construct-related validity findings. Exploratory factor analysis was performed on ratings within each exercise in order to determine the plausibility of the dimension activation theory versus the dominant dimension theory in distinguishing the cause of typical construct-related validity results. Subject matter experts (trained assessors) and the exercise designer were then asked to make judgments regarding the dimension activation potential/primacy of various dimensions of performance within the parameters of several assessment center exercises. These judgments were compared with each other and dominance analysis results obtained on ratings of dimensional performance in a series of managerial selection and development assessment centers. Specifically, within exercise dimension ratings were examined for importance with regard to the overall exercise ratings in the same exercise. Additionally, average rating variance estimates were obtained for dimensions deemed high in activation potential and compared with the average rating variance estimates for dimensions deemed low in activation potential.

Sample

The assessment center ratings were obtained on a sample of 97 managers participating in managerial selection and developmental assessment centers over a period of one year. Assessors consisted primarily of industrial/organizational psychology

graduate students working toward their Ph.D. at a large southeastern public university. Assessors had previously participated in a frame of reference based training program, consisting of a minimum of 24 hours of instruction and application feedback. In addition, all assessors had gained training experience observing and assisting with assessment in a private business setting. This business was also that from which part of the sample ratings were obtained. Additional data was collected as part of a leadership development program for an Executive and Physician's Executive MBA program at a large Southeastern university and based on the same assessment dimensions and exercises.

Sixteen assessors among those who initially made the dimension ratings were contacted to serve as subject matter experts with regard to the assessment center exercises and dimensions represented in the sample. These assessors' qualifications were based on their background in Industrial/Organizational Psychology (each had been involved in graduate study for a minimum of 2 years) and familiarity with the particular assessment center exercises and dimensions under evaluation (each had received substantial assessor training and participated in a minimum of 6 assessment centers engaging the relevant dimensions and exercises).

Assessment Center Description

The Assessment Center exercises and methods used in this study were developed and selected in response to the needs of two large organizations headquartered in the Southeastern United States and that of a large Southeastern university. The Task Force on Assessment Center Guideline's ethical and developmental standards served as a model for the creation of exercises and inclusion of dimensions. Between four and six exercises were used in the one-day assessment centers, with between five and fifteen dimensions

rated in each exercise. The number of exercises and dimensions analyzed were reduced based on the following criteria: a) dimensions must have been assessed in at least two exercises, b) dimensions and exercises had to be observed in a majority of cases in the study (Kudisch, Dobbins, & Ladd, 1997). Consequently, dimension ratings from four assessment center exercises were analyzed, including two simulation role-play exercises, an in-basket exercise, and a leaderless group discussion (Marsh & Grayson, 1995). The data included two dimension ratings of analysis, judgment, planning and organizing, decisiveness, leadership, delegation, initiative, coaching, team building, confrontation, sensitivity, and customer orientation, as well as an overall rating of exercise performance in each relevant exercise (see Appendix A for extended description of assessment center exercises and dimensions). As exercises were developed to elicit behaviors on a limited number of dimensions, only those dimensions relevant to the exercise were rated in each exercise (see Appendix B for a review of the dimensions assessed in each exercise). In particular, assessors observed, categorized, and documented candidate's behaviors with regard to relevant dimensions during the course of each exercise. Following this, assessors made separate dimension ratings based on the observations and recordings, as well as a final overall rating for exercise performance. Once all exercises were observed and dimensions were rated, assessors participated in a consensus discussion to obtain final dimension ratings. More specifically, each dimension was discussed and assessors shared relevant behavioral observations from appropriate dimensions to obtain the final ratings. This information was subsequently utilized in the creation of detailed feedback reports, prepared for consideration by the company of employ as well as individual candidates.

Dimension Activation Inventory

. The Dimension Activation Inventory is composed of a forced choice constant sum paired comparison of the comparative dominance or dimension activation potential of each dimension within each exercise (see Appendix C). Specifically, the direct estimation method was utilized in order to simplify subject matter expert judgments. Delivery of the Dimension Activation Inventory was via email and hard copy format. Participants were instructed as to the general purpose of the study and told that participation was voluntary and confidential, and completion of the Dimension Activation Inventory would serve as consent. Within the inventory, subject matter experts were provided with a definition of dimension activation potential and a paragraph outlining inventory instructions. More specifically, assessors were presented with all possible combinations of two dimensions within each exercise and asked to distribute a constant sum of points (100 points) within each pairing. Instructions indicated that the dimension having higher relative activation potential should receive a greater number of points. In addition to these items, demographic information was gathered for the subject matter experts and included questions regarding age, race, gender, education, years involved in personnel assessment, and number of assessment centers worked.

Design Indicator Scale

The design indicator scale was provided by the assessment center owner and developer and based on the exercise development strategy utilized in creating the exercises. More specifically, during creation, the developer had isolated dimensions of behavior to be the primary dimensions to be elicited in the contrived situation. The focus of exercise development centered around these dimensions, differing for each exercise.

Secondary dimensions were those likely to be elicited, but to a lesser degree. Tertiary dimensions were included as having the potential to be demonstrated, but less important to the overall performance within the exercise.

Statistical Analyses

Hypothesis 1: Exploratory Factor Analysis

In order to test hypothesis 1 and determine whether the dominant dimension model fit the data within each exercise or if it would be appropriate to consider a multi-dimensional model, two methods were considered. The first, confirmatory factor analysis, has frequently been utilized to compare competing models of exercise factor structure. As previously mentioned, Lance, Foster, Gentry, and Thoreson (2003) most recently compared models based on the salient dimension theory and a general impression theory using this method. However, while the salient dimension model used in this study could be accurately applied to a test of the dominant dimension theory in the current study, confirmatory factor analysis would not be conducive to modeling the structure of the dimension activation theory. Particularly, it would not be possible to predict just one appropriate pattern of factor loadings to model this theory. To further explain, consider an exercise in which two dimensions were labeled as activated and three dimensions non-activated. Within the dimension activation theory, it would be possible for each of the activated dimensions to load on separate factors, with non-dominant dimensions loading on yet another. That said, a model with both activated dimensions loading on a single factor and non-dominant dimension loading on a second factor would be acceptable, as well. Moreover, non-activated dimensions could conceivably have weaker loadings on factors representing dominant dimensions in

addition to a general non-activated dimension factor. Therefore, in order to avoid unnecessarily restricting the sample of possible factor patterns supporting the dimension activation theory, a second method, exploratory factor analysis, was considered. Using this method, it was possible to consider all possible models. Particularly, the number of factors extracted within each exercise could be specified based on expectations using the Maximum Likelihood extraction methodology. As an example, an exercise containing two activated and three non-activated dimensions could be examined with one, two, or three factors extracted.

Prior to conducting the exploratory factor analysis, the procedure necessitated that the structure of the rating data be simplified. Specifically, in conjunction with common assessment center method procedure, ratings from two assessors were available for each dimension within each exercise (the independent variables) and for overall exercise ratings (the dependent variable). In order to combine these judgments so that one rating could be referenced for each dimension as a data point in the factor analyses, the agreement of rater judgments ICC(A,2) was computed between assessor 1 and assessor 2 ratings (Shrout and Fleiss, 1979; McGraw and Wong, 1996). ICC(A,2) examines the variance between the two types of ratings with respect to residual variance, taking into account rank ordering as well as mean differences (McGraw and Wong, 1996). Once these estimates were obtained, the Spearman-Brown prophecy formula was applied in order to reveal the agreement level (Winer, 1962). Corrected intraclass correlation coefficients greater than .90 were taken as evidence of rating agreement (Bartko, 1976). Next, average ratings were computed between assessor ratings of within exercise dimension performance for each assessee.

Following this, exploratory factor analysis using maximum likelihood extraction and a varimax rotation was performed on the within exercise dimension ratings for each exercise. Three separate tests were run for each exercise specifying the extraction of one, two, or three factors. Significance testing served as an indicator as to the size of the second and/or third principle component and an indicator as to the appropriateness of the inclusion of more than one factor. Whenever an exercise reflects two or more activated dimensions, it is expected that the within exercise dimension ratings will represent more than one factor.

Hypothesis 2: Traditional Construct Related Validity Findings

In order to test Hypothesis 2, a multi-trait multi-method matrix was estimated to examine the mono-trait hetero-method and mono-method hetero-trait correlations. This method of correlational comparison was employed primarily to highlight the presence of dimension and/or exercise effects and to allow for the comparison of results with earlier studies. More specifically, in order to separate the average within exercise correlations from the average within dimension correlations, the mean correlation of each dimension rating with ratings of the same dimension in other exercises was computed for each dimension (mono-trait hetero-method). These mean dimension correlations were then averaged to form the average within dimension correlation for each group. Next, the mean correlation of each dimension rating with every other dimension rating within an exercise was computed for each exercise (mono-method hetero-trait). These mean exercise correlations were then averaged to form the average within exercise correlation. The average within dimension correlation was then compared with the average within exercise correlation using an independent sample t-test. Significant results were used as

an indicator that the average within dimension correlation differed from the average within exercise correlation. It was expected that differences would indeed exist with the average within exercise correlation exceeding the average within dimension correlation.

Hypothesis 3a: Aggregation of SME ratings of Dimension Activation Potential

In constructing the Dimension Activation Inventory, Comrey's (1950) methodology was chosen for use in determining relative dimension activation potential of various dimensions within exercises (Torgerson, 1958). Within this method, stimuli (dimensions) are presented as a series of paired comparisons, so that each dimension within an exercise is compared with each other dimension. Subjects were asked to distribute 100 points between each pairing based on judgments of the absolute ratio between them in terms of dimension activation potential (Comrey, 1950, as cited in Torgerson, 1958). This type of scaling procedure was selected as it represents a direct estimate of the ratios between each dimension pairing. Additionally, because the data are overdetermined in this case, a test determining it's nature as a ratio scale becomes unnecessary (Torgerson, 1958). In other words, consider n = the number of dimensions included in an exercise. Using this method, n(n-1)/2 ratios among dimensions will be considered and rated by subject matter experts as an indication of the scale value of each dimension within that exercise. As only n-1 ratios are necessary for determining scale values, a number of combinations of n-1 options are available for use. Considering each of the available options, an averaging procedure (arithmetic means) utilizes all of the obtained ratings and allows for an overdetermined solution.

As a test of hypothesis 3a, and in order to justify the aggregation of subject matter expert ratings of dimension activation potential, interrater agreement indices were

computed based on the results of the dimension activation inventory. To conduct this analysis, the sum of the points allotted to each dimension across parings was computed for each subject matter expert in each exercise separately. The interrater agreement of these dimension scores in each exercise was then assessed using the Rwg method. This method was chosen as it was developed for use in assessing agreement among a group of raters (James, Demaree, & Wolf, 1984; 1993). Within this method, the mean observed variance was compared to the expected variance based on the response format. Difference in observed and expected variance suggests a lack of agreement. Rwg estimates greater than .70 were considered indicative of a reasonable level of agreement (Burke, Finkelstein, & Dusig, 1999) and suggest the combination of subject matter expert ratings is appropriate.

Within each exercise, dimensions were consequently labeled as activated or non-activated within an exercise based on the dimension activation inventory results. Specifically, following the methodology proposed by Ladd, Atchley, & Burgess (2001) with regard to dominance analysis, dimensions were considered activated if the relative contribution of that dimension was greater than the mean contribution of all dimensions. Those dimensions in which the relative contribution was less then the mean contribution of all dimensions were considered to be non-activated in that particular exercise.

Hypothesis 3b: SME Ratings and Exercise Design Indicators

To assess similarity between SME (assessor) ratings of relative dimension activation potential and the purposeful design of the exercises, it was first necessary to determine the primacy of dimensions during the development process for each exercise. Establishing exercise design indicators required a consideration of the initial design

theory. When each exercise was created, the exercise designer focused on some dimensions as core dimensions. Other dimensions received secondary consideration, and a few received relatively smaller consideration. Therefore, data regarding the dimensions' relative consideration in exercise design was obtained via the exercise designer, and constituted a ranking of each dimension from one to three (primary to tertiary).

Following this, two types of correlational analysis were considered in attempt to discern the presence or absence of a relationship between the two scales for each exercise. First, a polyserial correlation coefficient was considered based on the nature of the two scales – a continuous activation potential scale and a trichotomous design indicator scale. A second coefficient, the phi coefficient, was also considered. This methodology requires the recoding of the data from each sample into dichotomous groupings representing activated/non-activated dimensions and design primary/non-primary dimensions. It must be noted that the low N of cases (ranging between eight and nine dimensions) renders both correlation coefficients to be unstable. In addition, the statistical power to detect relationships based on the small number of variables to be correlated is low in both cases. However, the phi coefficient was considered the most appropriate of the two analyses. In explanation, the design indicator scale by nature is not a ratio level scale. The primary dimensions represent those dimensions that were the focus of exercise design. Although a secondary set of dimensions existed in the design scheme, these dimensions were of considerable less focus. The distinction of most interest in this study is between the primarily activated/dominant dimensions and others. The phi coefficient considers these distinctions. Therefore, it was determined that the phi

coefficient should be utilized to examine all subsequent relationships when considering the design indicator scale. To this end, dimension scales were recoded to reflect their activated/non-activated status for the dimension activation potential scale and primary/non-primary status for the design indicators (with non-primary dimensions composed of secondary and tertiary dimensions as per exercise design specifications). A phi coefficient was computed between the recoded Dimension Activation Scale and the Design indicator scale for each exercise. High correlations were expected using the phi coefficient, with results significant at $p>.05$.

Hypothesis 4a: Dimension Dominance and Activation Potential

In order to test hypothesis 4a, agreement among dimension activation potential ratings and dimensional dominance in assessment center dimension ratings, it was first necessary to establish comparative dominance of the dimensions in each exercise. Several methods were considered for use in establishing the dominance of dimensions in each exercise as related to overall exercise performance, though Budescu's (1993) dominance analysis was ultimately determined to be most appropriate for this case. Mainly, traditional methods of estimating the relative contribution of variables to the dependent measure, such as standardized regression coefficients, zero-order correlations between a predictor and the dependent variable, and Darlington's (1968) usefulness statistic, fail to appropriately consider the case of correlated independent variables. Specifically, rather than accounting for a variable's independent effects as well as effects when combined with other variables, these methods tend to assign all shared variance to the strongest variable queried, leading to an exaggerated effect size for the strongest individual predictor and smaller than appropriate effect sizes for remaining variables

(Johnson, 2000). In other words, two variables may be highly correlated with each other and the dependent variable and yet be assigned very different regression weights. Moreover, as multiple regression implicitly implies that variables can be ordered in terms of their relative importance for prediction, calculations are made separately for each predictor and thus, relative importance is inferred from separate statistics (Budescu, 1993). Yet, predictors are often complexly interrelated. Conversely, two recent methods of importance analysis have addressed these concerns: Budescu's (1993) dominance analysis and Johnson's (2000) epsilon. Both methods allow for direct comparisons of relative importance in predictors. In particular, dominance analysis considers the average increase in Rsquared associated with a variable when considering that variable's direct effects, total effects, and partial effects (Budescu, 1993; Johnson, 2000). All possible combinations of variables are examined and as the sum of the variables' usefulness equals Rsquared, the relative weight of each variable can be computed by dividing it's estimated variance contribution into the total predicted variance when considering all variables. Within dominance analysis, when a dimension is the stronger predictor in all subset regressions, it is established as the dominant dimension. Conversely, Johnson's (2000) epsilon requires the creation of a variable set highly related to the original set, but uncorrelated. The relative weight of a predictor is calculated by dividing the proportion of variance in the new variable accounted for by the original variable by the proportion of variance in the dependent variable accounted for by the new variables (Johnson & LeBreton, 2002). As with Budescu's (1993) dominance analysis, the sum of the coefficients equal the model's squared multiple correlation, so the final weights can be

calculated by considering the proportional representation of each variable (Johnson, 2000).

Although both Budescu's (1993) dominance analysis and Johnson's (2000) epsilon were considered appropriate for said purposes, and have been shown to lead to only small differences in results (Johnson, 2000), Budescu's (1993) dominance analysis was chosen for three reasons. First, as it has been utilized for a longer period of time, it is generally a more well-known and thus accepted methodology. Second, the author deemed the method of analysis to be the more parsimonious of the two. Finally, this methodology shares similarities with the methodology used to determine dimension activation potential. Chiefly, both the dimension activation potential inventory and the dominance analysis method employed pairwise comparisons of each variable with each other variable.

Thus, following computation of dominance analysis, results were compared with dimension activation potential ratings in each exercise using two methods. Pearson's Product Moment Correlation Coefficients were computed between Dimension Activation Potential scores and Dominance Analysis results to determine consistency in dimension activation/dominance within each exercise. A significant coefficient greater than .70 will be indicative of a reasonable relationship (Cohen & Cohen, 1983). According to Hypothesis 4a, a strong correlation is expected.

Hypothesis 4b: Dimension Dominance and Design Primacy

As a test of hypothesis 4b, the phi coefficient was again utilized to examine the relationship between dimension dominance and exercise design indicators for each exercise. Accordingly, the Dominance Analysis results were first recoded to reflect their

dominant/non-dominant status. More specifically, in determining the relative dominance of one dimension over and above another dimension, paired comparisons were made between the contribution of each dimension in every possible model and sub-model. Next, a squared multiple correlation was computed based on these results, and a 95% confidence interval was specified for all pairwise differences (Budescu, 1993). The resulting grouping of dimensions in terms of their dominance was used in computing the Phi coefficient. More specifically, for three of the four exercises, four groupings of dimensions were revealed, with the first group being comprised of a solitary dimension. Thus, the first two of four groupings were labeled as dominant, with the last two given non-dominant status. For the fourth exercise (the In-Basket exercise) only two groupings of variables occurred. Therefore, the first was considered to be dominant to the second group. The bi-level Design Indicator scale separating the primary and non-primary dimensions in each exercise served as the comparison. Once more, a high correlation was expected with results significant at $p<.05$.

Hypotheses 5a: Activation Level Variance Comparisons

Variance estimates of assessment center ratings were computed separately for activated and non-activated dimensions. The average estimate for dimensions labeled as activated in an exercise was then compared with the variance estimate for dimensions labeled as non-activated. Comparisons among the activated and non-activated dimensions were made based on these estimates using Levene's Test for Equality of Variances. Specifically, Levene's test is the most common procedure for examining the equivalence of variance dispersions related to the variances within groups and is based primarily on discovering differences in the variability of the residuals (Hair, Anderson,

Tatham, & Black, 1995). Within this procedure, the absolute difference between a dimension rating and the mean rating for that dimension is computed. Following this, a one-way analysis of variance is conducted between the groups to determine whether the mean absolute deviation of the activated group differs significantly from the mean absolute deviation of the non-activated group (Neter, Kutner, Nachtsheim & Wasserman, 1996). Significant differences in variance were expected at $p<.05$, with the average activated dimension having higher variance than the average non-activated dimensions in each exercise.

Hypothesis 5b: Dimension Primacy Variance Comparisons

For each exercise, dimensions were first split into three groups consistent with the exercise design scale. Variance estimates for assessment center dimension ratings were computed separately for each group with Group 1 representing those dimensions for which the exercise was primarily designed to represent, Group 2 representing the secondary dimensions, and Group 3 representing the least prominent dimensions. It was expected that the average variance estimate for Group 1 dimensions would exceed the average variance estimate of Group 2 and Group 3, and the average variance estimate of Group 2 would exceed the average variance estimate of Group 3. The equality of these estimates was compared using Levene's Test of Equality of Variances with differences expected to be significant at the $p<.05$ level.

CHAPTER IV

RESULTS

A Consideration of Hypotheses

Hypothesis 1: Exploratory Factor Analysis

In order to simplify the structure of the assessment center data by combining the two rater judgments for each within exercise dimension ratings, it was first necessary to determine the level of rating agreement between said raters. Within each of the four exercises examined, ICC(2) was computed between rater 1 and rater 2 judgments. Following the application of the Spearman-Brown prophecy formula, it was found that within the In-Basket exercise and Role Play 2 there were very high levels of agreement (e.g., .90 and .91, respectively). While results for the Group Decision Making exercise and Role Play 1 were less strong, ratings nevertheless showed strong levels of agreement (e.g., .82 and .87, respectively). Thus, the decision was made to combine separate rater judgments into mean scores using an averaging procedure.

For each exercise separately, three exploratory factor analyses were performed using a maximum likelihood extraction method specifying one, two, and three factors. In one instance, the In-Basket exercise, the presence of Haywood cases (communalities greater than one) created considerable problems in the estimation of models, resulting in non-positive definite solutions. Therefore, corrective action was taken by fixing the communalities of those dimensions at 1.0. According to Hypothesis 1, it was expected that more than one factor would be necessary to best represent the dimension ratings for each exercise. As shown in Table 1, consistent with expectations, two factors were

Table 1. Exploratory Factor Analyses

Exercise	*Factors*	*Chi-square*	*Df*	*Significance*
Group Decision Making Exercise	1	52.817	20	.000
	2	23.952	13	.032
	3	4.965	7	.664
Role Play 1	1	38.354	27	.072
	2	23.337	19	.223
	3	N/A	N/A	N/A
Role Play 2	1	65.040	27	.000
	2	38.871	19	.011
	3	18.280	12	.107
In-Basket Exercise	1	1577.77	20	.000
	2	810.72	13	.000
	3	53.45	7	.000

necessary for representation of ratings in both the Group Decision Making exercise and Role Play 2 (e.g., two factor models significant at $p<.032$ and .011, respectively), and three or more factors were necessary in the case of the In-Basket exercise (three factor model significant at $p<.001$). However, for Role Play 1, one factor sufficiently described the data. Thus, for three of the four exercises, the hypothesis was supported.

Hypothesis 2: MTMM Analysis

In order to directly compare the results of the present study with results from published construct-related validity examinations, a multi-trait multi-method framework was used to reveal average within dimension and within exercise rating correlations. In particular, the convergence of dimension ratings was examined by computing average within dimension rating correlations. Conversely, as represented by the average within exercise rating correlations, evidence for the discrimination of dimensions was examined. According to Hypothesis 2, and in harmony with related past studies, it was expected that the average within exercise rating correlations would exceed the average within dimension rating correlations. Demonstration of this rating pattern would traditionally be interpreted as a lack of convergence and discrimination (and thus a lack of construct-related validity) among relevant dimensions. Consistent with expectations, the average within exercise rating correlation (mean $\underline{r} = .43$, $\underline{SD} = .17$) exceeded the average within dimension rating correlation (mean $\underline{r} = .18$, $\underline{SD} = .14$) (see Table 2), and these differences were statistically significant ($\underline{t} = -8.338$, $\underline{p}<.001$). Moreover, the average within dimension rating correlation was nearly equal to the average correlation between ratings of different dimensions in different exercises (hetero-dimension hetero-exercise correlation) (mean $\underline{r} = .14$, $\underline{SD} = .11$). Thus, hypothesis 2 was supported.

Table 2. Descriptive Statistics for Mean Within Dimension and Within Exercise Rating Correlations

	N	*Range*	*Mean r*	*SD*
Within Exercise	128	-.22 to .914	.4301	.1740
Within Dimension	33	.008 to .791	.1788	.1441
Heterodimension-Heteroexercise	390	.001 to .533	.1362	.1075

Hypothesis 3a: Aggregation of Subject Matter Expert Ratings

The Dimension Activation Inventory was delivered to 16 subject matter experts in both hard copy format as well as email format. Twelve participants completed all portions of the inventory (75% response rate); however, one set of responses was excluded from analyses due to a clear misunderstanding of the directive as the individual did not discriminate among any of the dimensions. Thus, dimension activation potential scores were ultimately based on the judgments of 11 subject matter expert ratings (descriptive sample details can be found in Table 3). The subject matter experts were primarily female (92%), with a mean age of 28.5 (SD = 3.38), and had rated on average 54 assessment centers (SD = 45.69) and 242 candidates (SD = 120.90). Each participant was asked to read a brief definition of Dimension Activation Potential and instructed to distribute 100 points among each dimension pairing. This distribution was to be based on judgments regarding the relative activation potential of each dimension. The process was repeated for each of four exercises separately.

Agreement among dimension activation potential ratings was investigated using Rwg statistics. Agreement among raters was universally high, and ranged between Rwg = .9782 for Role Play 2 and Rwg = .9853 for the Group Decision Making exercise. Therefore, hypothesis 3a was supported, justifying aggregation of the subject matter expert ratings into a ratio scale.

Next, using the method proposed by Comrey (1950) to develop a ratio level scale, scale scores were computed independently for each exercise. The mean dimension scale values were .57 for the In-Basket exercise, .62 for Role Play 1, .67 for Role Play 2, and .78 for the Group Decision Making exercise. Ultimately, within each exercise, all

Table 3. Mean Statistics for Subject Matter Expert Sample

N	*Age*	*Assessment Experience*	*Assessment Centers Rated*	*Candidates Rated*
11	28.36	5 years	51	241.8

dimensions with scale values exceeding the mean for that exercise were labeled as activated dimensions, whereas all dimensions with scale values not exceeding the mean value were labeled as non-activated for that exercise (see Table 4). To summarize, the dimensions of Analysis and Judgment were found to be activated in each exercise included in this study. Furthermore, the dimension of Leadership was activated in each of the interpersonal exercises (it was not rated in the In-Basket exercise). That said, each of the interpersonal exercises included one dimension that was not activated in any other exercise. More specifically, Role Play 1 included Confrontation as an activated dimension, as well as Decisiveness (which was common with the In-Basket exercise). Conversely, Coaching skills were activated in Role Play 2, and Team Building was activated in the Group Decision Making exercise. However, no activated dimension rated in the In-Basket exercise was unique only to that exercise, although scale value of the dimension of Initiative nearly met the criteria (scale value = .55).

Hypothesis 3b: SME Ratings and Exercise Design Indicators

It was hypothesized that SME ratings of dimension activation potential and details of the primacy of dimensions intended during exercise design would closely correspond. Due to the nature of the initial design model and the low N of cases included in this analysis, a phi correlation coefficient was utilized to assess congruence. With the recoding of the data to represent activated/non-activated dimensions for the Dimension Activation Potential Scales and primary/non-primary based on Design Indicators, the phi coefficient was computed to determine the relationship between the two scales for each

Table 4. Dimension Activation Inventory Scales

Exercise	*Oral Communication*	*Analysis*	*Judgment*	*Planning Organizing*	*Decisiveness*	*Delegation*	*Leadership*	*Initiative*	*Coaching*	*Team Building*	*Confrontation*	*Sensitivity*	*Customer Orientation*
In-Basket		**.95**	**1**	.47	**.69**	.32		.55		.18			.43
Role Play 1	.28	**.82**	**1**		**.68**	.4	**.89**		.47		**.67**	.38	
Role Play 2	.31	**.99**	**1**		.6		**.83**	.54	**.74**		.54	.46	
Group	.43	**.89**	**1**				**1.17**	.74		**.82**	.65	.56	

Note: All Activated Dimension are in boldface.

exercise. To summarize, a significant correlation was found between the dimension activation potential indicators and design indicators for Role Play 2 (phi = .756, $p<.023$), but not for the In-Basket exercise, Role Play 1, or the Group Decision Making exercise. That said, correlations between the two indicators were of sizable note for Role Play 1 ($r = .632$) and the Group Decision Making exercise ($r = .577$) despite the small number of variables. Therefore, it appears there may be a relationship between dimension activation potential ratings and the initial design strategy for three of the four exercises (See Table 5 for summary).

Hypothesis 4a: Dimension Dominance and Activation Potential

In computing the dominance analysis, it was necessary to eliminate dimensions in two of the exercises. In explanation, within the assessment center rating data, there were a significant number of instances in which an individual did not receive a rating for a dimension despite its inclusion in the exercise. Most typically, the missing data point resulted from insufficient behaviors to warrant a formal rating. When a missing data point is encountered during the dominance analysis procedure, the relevant subject is entirely eliminated from the analyses. Due to a high number of missing data points for the dimensions of coaching and delegation in Role Play 1, and the dimension of confrontation Role Play 2, it was necessary to eliminate these dimensions during the dominance analysis procedure. Inclusion of the dimensions would have significantly altered the sample size utilized in the analysis. As none of these dimensions were labeled as activated according to SME ratings or primary according to design indicators, elimination of the dimensions was considered reasonable. As a result, the number of dimensions included in the dominance analysis for each exercise was as follows:

Table 5. Correlations Between SME Dimension Activation Potential Rating and Design Primacy Indicators

Exercise	*Phi Coefficient*	*p-value*
In-Basket	.149	.673
Role Play 1	.632	.058
Role Play 2	.756	.023
Group	.577	.102

In-Basket exercise (n=8), Role Play 1 (n=7), Role Play 2 (n=7), Group Decision Making exercise (n=8).

Table 6 outlines the results of the dominance analysis for each of the four exercises. Using Pearson's Product Moment correlation coefficients, dimension importance indicators were significantly correlated with ratings of dimension activation potential for both Role Play 1 ($r = .791$, $p<.034$) and the Group Decision Making exercise ($r = .806$, $p<.016$). However, these same correlations were non-significant for the In-Basket exercise ($r = .689$, $p<.059$) and Role Play 2 ($r = .311$, $p<.453$) (see Table 7). In sum, Hypothesis 3b was only supported for the exercise Role Play 2 and the Group Decision Making exercise. However, in interpreting these results, it must be noted that the presence of sizeable correlations despite a clear lack of statistical power to detect the relationships suggest stronger conclusions than can be made based on significance testing. In particular, with the power for detecting a correlation of .70 in the In-Basket exercise at less than .60, it is reasonable to conclude that a relationship does exist in this case despite the lack of significance.

Hypothesis 4b: Dimension Dominance and Design Indicators

Hypothesis 4b anticipated a strong correlation between the primacy of dimensions as originally intended in exercise design and the importance of dimension ratings in predicting overall exercise performance ratings. However, contrary to expectations, correlations between design primacy indicators and dimension relative importance were low to moderate and non-significant for all but one exercise as per phi correlational analyses. Phi coefficients based on recoded data revealed only one significant relationship for the exercise Role Play 2 (phi = .745, $p<.035$) (See Table 8 for summary).

Table 6. Dominance Analysis Relative Importance Scores

Exercise	*Oral Communication*	*Analysis*	*Judgment*	*Planning Organizing*	*Decisiveness*	*Delegation*	*Leadership*	*Initiative*	*Coaching*	*Team Building*	*Confrontation*	*Sensitivity*	*Customer Orientation*
In-Basket		.23	.22	.04	.15	.15	.02			.08			.12
Role Play 1	.04	.16	.29		.11			.09			.13	.18	
Role Play 2	.08	.05	.24		.03		.19	.19			.12	.12	
Group	.04	.18	.21				.17	.17		.11	.03	.09	

Table 7. Correlations Between SME Dimension Activation Potential Ratings and Dimension Importance via Dominance Analysis

Exercise	*Pearson Correlation*	*p-value*
In-Basket	.689	.059
Role Play 1	.791	.034
Role Play 2	.311	.453
Group Decision	.806	.016

Table 8. Correlations Between Dimension Importance via Dominance Analysis and Design Primacy Indicators

Exercise	*Phi Coefficient*	*p-value*
In-Basket	.333	.346
Role Play 1	.091	.809
Role Play 2	.745	.035
Group Decision	.447	.206

Hypotheses 5a: Activation Level Variance Comparisons

As a test of Hypothesis 5a and in order to compare the difference among the activated dimension variances and the non-activated variances, dimension ratings within each of the exercises were separated into two groups, depending on their activation status. Levene's test for the Equality of Variances was then conducted, with variances expected to be non-equal. Results are outlined in Table 9. In support of Hypothesis 5a, highly significant differences in were found for each exercise, with the variance estimates of the activated dimensions exceeding the variance estimates of the non-activated dimensions in each case. Specifically, the variance estimates for the activated dimensions in the In-Basket exercise, Role Play 1, Role Play 2, and the Group Decision Making exercise were .4176, .2896, .3846, and .2771, respectively. Variance estimates for the non-activated dimensions were .2441, .1983, .2302, and .1851, respectively. Differences were significant at $p<.001$.

Hypothesis 5b: Dimension Primacy Variance Comparisons

As a test of Hypothesis 5b, dimension ratings were separated into three groups, depending on the primacy of the dimension during the exercise design initiative, with Group 1 representing primary dimensions, Group 2 representing secondary dimensions, and Group 3 representing tertiary dimensions. Levene's test for Equality of Variances was conducted between each of the groups for each exercise. Results are outlined in Table 10. In partial support for Hypothesis 5b, variance estimates for primary dimensions exceeded variance estimates for tertiary dimensions for each of the exercises. However, excluding a significant difference between variance estimates for secondary versus tertiary dimensions in the In-Basket exercise ($f = 15.48$, $p<.000$), no other variance

Table 9. Levene's Test for Activated Versus Non-activated Dimensions

Exercise	*Acitvated Dimension Variance*	*Non-activated Dimension Variance*	*F statistic*	*P value*
IB	.4176	.2441	23.97	.000
RP1	.2896	.1983	21.77	.000
RP2	.3846	.2302	44.28	.000
GDM	.2771	.1851	19.33	.000

Table 10. Levene's Test Based on Exercise Design Primacy

Exercise	*Primary Dimension Variance*	*Secondary Dimension Variance*	*Tertiary Dimension Variance*	*F test for Primary vs. Secondary*	*p-value*	*F-test for Primary vs. - Tertiary*	*p-value*	*F-test for Secondary vs. -Tertiary*	*p-value*
IB	.3703	.3399	.2318	.00	.960	10.77	.001	15.48	.000
RP1	.2788	.2490	.2112	.43	.513	5.65	.018	3.07	.080
RP2	.3535	.2830	.2638	3.35	.067	9.73	.002	2.12	.145
GDM	.2829	.2150	.1966	2.94	.087	5.17	.023	.78	.378

comparisons revealed significant results. Therefore, overall, Hypothesis 5b was partially supported.

CHAPTER V

DISCUSSION

The demonstration of construct-related validity in assessment center ratings has been a focal point of the academic literature for decades. The majority of such studies have focused their evaluations on the application of Campbell and Fisk's (1959) multi-trait multi-method framework to assessment center ratings. In particular, within this method, dimension performance ratings within an exercise are expected to have lower correlations than ratings of the same dimension across exercises. However, typical findings have revealed the opposite pattern of rating correlations, with those dimensions within an exercise often evidencing significantly stronger relationships. As a result, numerous researchers have concluded that dimensions are not construct valid in that raters are unable to discriminate among dimensions when rating within exercise dimension performance. The primary purpose of this study was to review expectations for construct-related validity within assessment centers in light of exercise design specifications, and to establish a new perspective with which to view multi-trait multi-method results. To that end, two competing theories that have been discussed to explain high within exercise rating correlations were examined. The first, the dominant/salient dimension theory, has suggested that the high intercorrelations among dimension ratings within exercises result from halo. A second theory that has more recently been applied to the assessment center construct validity issue is the trait activation theory. Originating in the personality literature, the trait activation theory suggests the pattern of correlations found in assessment center dimension ratings are a result of different traits being

activated in different exercises. As an extension of this theory, the dimension activation theory suggests that different dimensions are activated in different exercises, causing reduced cross-situational consistency in ratings based on different levels of activation. In addition, different levels of activation among dimensions within exercises should also result in different levels of rating variance. Specifically, low levels of variance in non-activated dimensions will ultimately cause high levels of rating intercorrelation. It was expected that the pattern of activation among dimensions could be established by consulting the exercise design strategy and beliefs by subject matter experts regarding their relative activation potential.

The results of the study and the implications of findings will first be discussed within this chapter. Following this, study limitations will be addressed. Lastly, the chapter will conclude with a detailed consideration of future research agendas related to the findings.

A Traditional Examination of Construct Related Validity

In order to provide a case for the generalizability of the findings in this study, evidence of the similarity of rating data structure was sought. As mentioned above, much of the research examining assessment center construct related validity has been framed in terms of Campbell's and Fisk's (1959) multi-trait multi-method framework. Therefore, the same procedure was employed in this study in order to provide a comparison. An examination of construct related validity results of dimension ratings based on the multi-trait multi-method analysis show that the pattern of intercorrelations found in the current data is similar to that found in previous studies. More specifically, as is common in such studies, the within exercise rating correlations significantly exceeded the within

dimension rating correlations. This pattern stands in direct contrast to Campbell and Fisk's (1959) criteria for demonstrating convergent and discriminant validity in the ratings.

A common interpretation of these results would suggest that the data is not construct valid. In other words, ratings do not actually represent behavior on performance dimensions included in the exercises. However, one important point that is commonly overlooked in assessment center construct validity research must be highlighted on this occasion. That is, assessment center exercises actually represent a common method. As application of the multi-trait multi-method framework would require the treatment of different exercises as different methods, the application of this framework is clearly inappropriate.

A consideration of expectations based on the Dimension Activation Theory further recommends a suspension on the interpretation of the multi-trait multi-method results. It was the purpose of this study to attempt to understand the source of the high within exercise rating correlations as traditionally and presently found in assessment center rating data. As theorized, it was expected that these correlations were specifically a result of high correlations among the ratings of non-activated dimensions. Resulting from lower levels of rating variance, these highly correlated dimensions are a function of exercise design strategy, can be predicted, and should be expected. The following hypotheses served as a check for this theory.

Dimension Activation Theory vs Dominant Dimension Theory

Two of the hypothesis were directly relevant for establishing a foundation for the examination of the Dimension Activation Theory. First, the Dimension Activation

Theory could not be realistically expected to hold explanatory power in the case of pervasive halo within exercises. Therefore, in order to discredit the competing Dominant/Salient Dimension theory, exploratory factor analyses were conducted. In opposition to the latter theory, more than one factor was necessary to best represent the data in three out of the four exercises. Role Play 1 was the sole exercise in which one factor was sufficient to describe the data. Interestingly, this exercise was also the only exercise in which more than half of the dimensions were labeled as activated. Moreover, an examination of factor loadings of the various dimensions included in this exercise revealed that the strength of the dimension loadings could be predicted based on the activation potential of the ratings within this exercise. Specifically, the dimension's potential activation level (as determined by subject matter expert ratings of dimension activation potential) was significantly correlated with the dimensions loading on the one factor ($r = .785$, $p < .012$). Thus, although for this one instance the evidence did not suggest that the dominant/salient dimension theory was inappropriate, neither did the evidence conclusively disconfirm the dimension activation potential theory.

Supplementary evidence in support of a consideration of the Dimension Activation Theory emerged from an examination of the Dimension Activation Inventory ratings. In particular, a high level of agreement regarding the relative activation potential of dimensions within exercises was found among the eleven raters serving as subject matter experts. Moreover, there clearly existed differentiation among some dimensions in terms of activation potential ratings. Taken together, high convergence of ratings and variability in dimension activation levels within a given exercise suggests that dimension

activation is a concept worthy of examination. It further suggests that the equal treatment of all dimensions included in any given exercise is inappropriate.

Activation as a Function of Design

The fact that agreement exists regarding relative activation potential is a particularly evocative finding as this represents the first study to consider differential dimension activation levels. Of further interest, however, was the finding that subject matter expert ratings of relative activation potential were not correlated with the manner in which the exercises were designed for three of the four exercises. A cursory examination of these results would suggest that the intended design of the exercises – to elicit a few dimensions of behavior in particular – failed. However, closer inspection of the results paints a very different picture. Particularly, for all but the dimension of planning and organizing in the In-Basket exercise and the dimension of Initiative in Role Play 2, every dimension that was primary in the initial exercise design was also rated as activated by subject matter experts. Moreover, although these two dimensions were not labeled as activated, they were not far behind. Therefore, it is apparent that the exercise design strategy was successful in its primary purposes. For those dimensions that were of lesser importance when creating the exercise situations, there was less agreement regarding activation levels. In some cases, ample behaviors relating to some dimensions were evoked despite the fact that the exercise was not designed primarily for that purpose. In other cases, dimensional performance deemed as having secondary potential by the exercise designer did not elicit strong behaviors as per subject matter experts. Furthermore, it was these minor differences that affected correlational results. Hence, it can be concluded that some dimensions are more likely to be activated than others within

an exercise. Moreover, most primary dimensions are activated because the exercise situations were designed for that purpose. These findings deserve special note. While several past studies have noted the dominance of one dimension or another in the rating process, this is the first study to actively consider the source of the differing levels of dominance. It appears that the answer to this question may even be comically simple – because the exercises were designed that way.

Relationship Between Activation Level and Dominance

A consideration of the role of dimension activation in predicting overall exercise ratings was similarly encouraging. Performance on activated dimensions appeared to be used when judging overall exercise performance in some exercises (as per dominance analysis), though activation level did not appear to be a factor universally. Particularly, correlations based on Role Play 1 and the Group Decision Making exercise fostered significant results. However, non-significant results were obtained for the In-Basket exercise and Role Play 2, the latter for which a very low and non-significant correlation was found. It must be noted that for Role Play 2, it was necessary to eliminate one dimension from the dominance analysis to deal with missing data points. Thus, it is conceivable that the inclusion of this dimension could have altered results obtained. That said, two dimensions were eliminated from the dominance analysis for Role Play 1, as well, and a significant relationship was found between dimension activation rankings and dominance analysis results in this exercise.

Despite this, the overall findings linking the activation indicators to dominance analysis results were quite encouraging. In the case of the In-Basket exercise, non-significant results were nevertheless based on a moderate correlation (r = .69). In light

of the low level of power estimated for detecting these relationships, findings indicate that the relationship between these indicators may approach significance with a larger sample size. In summary, it appears as though raters may rely on performance on activated dimensions as primary when making assessments of overall exercise performance, though results did not hold for one exercise.

Somewhat in contrast, when examining the correspondence between design indicators of dimension importance and dominance analysis results, low and non-significant correlations were found between the scales for all but one exercise, Role Play 2. Once again, this exercise had previously been found to be the only exercise in which non-significant results were found when comparing activation potential and importance via dominance analysis. It must be noted, however, that this is also the exercise in which the dimension of Initiative was primary in the design procedure, but did not meet the criteria to be labeled as activated by subject matter experts.

At this point it is imperative to call attention to one of the major limitations of this study. As the number of dimensions included in the four exercises for the dominance analyses ranged between seven and nine, the sample sizes used in conducting the correlational analyses were less than ideal. Specifically, the low number of variables included in these procedures resulted in low levels of power for detecting relationships between the scales. However, the detection of even moderate, non-significant correlations when relying on power levels at or below the .50 level suggests the presence of relationships among the variables. Altogether, though, on the basis of the intercorrelations found in the present study, it must be concluded that the neither the

activation potential nor the primacy of dimensions within an exercise can be conclusively relied upon as the performance indicators used in determining overall exercise ratings.

An Examination of Variances

An examination of the difference in variance estimates between activated and non-activated dimensions revealed that in accordance with expectations set forth in the Dimension Activation Theory, activated dimensions have significantly more variance in ratings between individuals than non-activated dimensions. This was true for all four exercises included in the study. Furthermore, when all dimensions were considered separately, each activated dimension showed greater variance than each non-activated dimension in the In-Basket exercise. Whereas the non-activated dimension of initiative showed higher levels of rating variance than one or more activated dimension in the Group Decision Making exercise and Role Play 2, this was not detrimental to supporting the theory. Particularly, in the Group Decision Making exercise, initiative was very nearly classified as activated. More specifically, the initiative activation rating was .74 and the cut-point for activation classification was .78. On the other hand, in Role Play 2, the dimension of initiative was labeled as primary by the design indicator scale. Moreover, though not missing the cut-point at nearly as close of a range as reported above, initiative was the dimension closest to being included in the activated group for this exercise. Take together, it may be that initiative is indeed an activated dimension for this exercise; however, the subject matter experts did not key into its importance to the exercise.

Nevertheless, there was one exercise for which the results were less clear. In Role Play 1, the variance estimate for one of the activated dimensions (e.g., confrontation) was

lower than variance estimates of three non-activated dimensions (e.g., delegation, coaching and sensitivity). For a separate non-activated dimension (e.g., analysis), the variance estimate exceeded that of one activated dimension (e.g., delegation). However, when taken as a whole, activated dimensions still had significantly larger variance estimates than non-activated dimensions. See Appendix D for a table highlighting the variance estimates for each dimension in each exercise.

When variance estimates were examined based on the primacy of dimensions in exercise design, there were significant differences between primary and tertiary dimensions with primary dimensions having higher levels of rating variance than tertiary dimensions. In one instance, the In-Basket exercise, secondary and tertiary dimension rating variances significantly differed. That said, secondary dimension variances were distinguishable from no other variances (primary or tertiary) in any other comparison. As a post hoc procedure, the same analyses were performed between primary dimension ratings and that of all other dimensions (non-primary dimensions) based on design indicators. Results revealed significant differences between primary and other dimensions for two of the exercises (e.g., Role Play 2 and the Group Decision Making exercise), but non-significant results were found in the other two incidences.

Summary and Implications

Exploratory factor analyses suggest that the dominant/salient dimension (halo) theory does not hold up for most of these exercises. Moreover, subject matter expert ratings of activation potential corresponded well with variance differences. In other words, those dimensions that are more likely to be activated generally show higher levels of rating variance. In addition, the non-activated dimension variances are quite small.

Taken together, this evidence lends support to the contention that low levels of rating variance among non-activated dimensions could be the source of the high within exercise rating correlations. Given the evidence that most of the activated dimensions are expected to elicit more relevant behavior by exercise design, negative conclusions for the construct related validity of assessment center ratings based on the multi-trait multi-method framework should be questioned. In contrast, the evidence presented in the study suggests that exercises are designed to elicit some dimensions of behaviors more so than others. Additionally, subject matter experts familiar with the exercises and dimensions show high levels of agreement regarding which dimensions are most likely to be activated in a given situation. In this study, less than half of the dimensions included in three of the four exercises were activated. For those exercises, non-activated dimensions almost unilaterally generated lower levels of rating variance than activated dimensions. Thus, significant levels of intercorrelations among dimension ratings within an exercise should be expected. In light of this, the multi-trait multi-method framework and interpretation directives may be inappropriate for assessing the construct related validity of assessment center ratings.

Taken together, the evidence strongly indicates that criteria for demonstrating construct-related validity of assessment center ratings needs revision. Specifically, the nature of exercises and design strategy should be considered when setting forth expectations regarding the intercorrelation of dimension ratings. Moreover, it must be acknowledged that previous findings do not negate the potential of construct valid ratings. In fact, this study represents an alternative manner in which to assess the construct related validity of assessment center ratings and supports the validity in most

cases. Primarily, subject matter experts confirmed the activation of most dimensions that the exercises were designed to elicit. Moreover, those dimensions that were activated also showed higher levels of rating variance further supporting their activation. Though the non-activated dimensions evidence lower levels of discriminate validity, the lack of variance is expected and thus, does not contradict the assertion that ratings actually represent performance on those dimensions. Nevertheless, the ratings of these non-activated dimensions are expected to be less important to the assessment procedure for that particular exercise and thus, should not be assigned as great of levels of importance.

Methodological Limitations

Noted above, one of the primary limitations of this study resides in the fact that only eight to nine dimensions were included in each of the exercises. This low number of variables significantly limited the power of correlational analyses in detecting relationships between dimension activation, design indicators, and dimension importance. As a result, some relationships that might be significant were found to be non-significant, limiting support of the theory. Nevertheless, the inclusion of a significantly larger number of dimensions in the exercises would be detrimental to the process. Specifically, in a study conducted in 1989, Gaugler & Thornton found that increasing the number of dimensions included in an exercise negatively affected the accuracy of dimension ratings. Additional research has likewise suggested and demonstrated that limiting the number of dimensions observed in an exercise improved construct validity and resulted in higher proportions of rating variance (Arthur, Woehr, & Maldagen, 2000; Lievens & Conway, 2001; Woehr & Arthur, 2003).

The sample included in this study serves as yet another methodological limitation. In order to obtain a sample for which subject matter experts were familiar with the exercises and dimensions rated, and obtain overall exercise performance ratings to be used in the dominance analysis, the sample size was limited to ninety-seven subjects on which ratings were gathered. While it was determined that this sample size was sufficient for the study purposes, a larger sample would have been preferable to enhance the validity of the results.

Third, it must be mentioned that the necessary familiarity of the subject matter experts with the exercises and dimensions caused a potential confounding effect with the assessment center ratings. Many of the subject matter experts were the same individuals that served as role players and raters in the assessment center rating data set. Moreover, each of these experts/raters had necessarily participated in training sessions prior to becoming assessors/role players. During the course of this training, instruction included detail regarding the relative importance of some dimensions in the various exercises. Specifically, assessors were directed as to the necessity of rating the primary dimensions as per exercise design. Hence, a bias may have existed in the rating of the relative activation potential of various dimensions across exercises that resulted in agreement among design indicators and activation potential ratings for those primary dimensions. Nevertheless, in order to reduce those specific training effects, effort was made to obtain only subject matter experts who had rated a large number of assessment centers and had done so over a significant period of time. As intensive training was conducted prior to employment for the Assessment Center, a considerable period of time elapsed between instruction and the activation potential rating process. Nevertheless, confounding effects

cannot be ruled out. That said, this instruction should not have had a significant effect on rating variance results. As rating variance results supported the validity of the Dimension Activation Theory, the potential for confounding does not meaningfully affect the overall conclusions.

Exercise Limitations

One of the major limitations of any assessment center construct related validity study is one of generalizability of results to other assessment center exercises. As expected, the exercises utilized in this study comprised two role play situations, a group decision making task, and an in-basket exercise. Although all of these formats are common to assessment centers, the specific situations and focal dimensions vary from assessment center to assessment center. In addition, the original design strategy for these exercises may or may not be unique to this assessment center. This highlights a glaring weakness in assessment center guidelines. Principally, there currently exists no standard methodology for the development of exercises. This lack of standardization will continue to limit the applicability of study results. Furthermore, the lack of discussion in the academic literature regarding exercise design strategies highlights the lack of consideration of design specification. It begs the question, "Are exercises being designed in the hopes of equally eliciting performance on a large number of dimensions? Is this possible? Is this practical?" It the case of the current study, the greater the number of dimensions that were activated, the more problems there were that occurred with demonstrating construct-related validity. More specifically, in the case of Role Play 1, the differences in variance estimates between activated and non-activated dimensions were not as clean as for other exercises with fewer dimensions activated. Moreover,

when comparing activation to importance categorizations, this was the only exercise for which results were non-significant and could not be reconciled. Lastly, this was the only dimension that needed only one factor to best represent the data via exploratory factor analyses.

Future Research

A Replication of Findings

As a preliminary direction for future research, and in order to replicate present findings, consideration must be given to the design strategies currently used in the creation of assessment center exercises. As mentioned above, the design strategy utilized in this study is by no means suggested to be representative of the typical approach as currently, no typical procedures have been reviewed or suggested. Therefore, it is of primary import to discern an understanding of exercise design strategies and test out the Dimension Activation Theory on similar as well as dissimilar approaches. Specifically, whereas a number of studies have focused on discerning the effects of exercise content (competitive vs. cooperative) and form (role play vs. group discussion), prior to this study, no such focus has been placed on exercise purpose (Schneider & Schmitt, 1992). This replication of findings would serve to reveal the generalizability of results. Moreover, as design strategies are clarified and categorized based on similar features, the next step will be to determine the relative construct related validity obtained using each style, as well as content and criterion related validity. According to the unitarian framework of validity (Binning and Barrett, 1989), it is expected that the methodology that supports construct-related validity should be the same that supports content and criterion-related validity. With this in mind, construct-related validity evidence should be

obtained in a manner consistent with the design strategy. More specifically, according to the Dimension Activation Theory, those dimensions that a given exercise was designed to primarily elicit should show greater levels of rating variance than the dimensions of lesser import.

These findings should also be replicated with the use of subject matter experts not trained in the specifics of the exercise design strategy, though trained in the assessment center methodology. More specifically, in the case that subject matter experts are not provided with any details about the relative importance of the various dimensions included in an exercise, confounding effects of training will be eliminated. Ability to ascertain the relative activation potential of the relevant dimensions would further support the Dimension Activation Theory.

Integration of Information

With further support of the Dimension Activation Theory, a necessary next step is to ascertain the impact of the ratings of activated versus non-activated dimensions on the overall assessment ratings. This is particularly important because despite its purported consequence, the information integration process has largely been ignored in the assessment center academic literature (Lievens & Klimoski, 2001; Zedeck, 1986). Though in the current instance, it was revealed that assessors do not reliably utilize performance information in the activated dimensions above the non-activated dimensions in judging overall exercise performance, it is unclear as to the impact that performance on activated dimensions has on overall assessment ratings. Of note, in this study, overall exercise ratings were collected purely for research purposes with the understanding that no decisions would be made based on the ratings. However, overall assessment ratings

are often used to make critical decisions and recommendations. Therefore, it is expected that ratings of activated dimension/exercise combinations would potentially have stronger impact on overall assessment ratings than non-activated dimensions. An examination of the effect of such a decision strategy (versus contrasting decision strategies) on the criterion-related validity of overall assessment ratings should follow.

Alternative Explanations

Several recent academic articles have directly applied the trait activation theory to the assessment center construct-related validity research (Tett, 1998, 1999; Haaland & Christiansen, 2001). Specifically, traits have been conceptualized as "intraindividual consistency and interindividual uniqueness in propensities to behave in identifiable ways in light of situational demands" (Tett & Schleicher, 2001). As exemplified by Tett & Schleicher (2001), traits are distinguished from assessment center dimensions in that dimensions are inherently valued, whereas the value of a particular trait is dependent on the situational context. Additionally, traits have greater psychological depth.

All in all, there appears to be some initial evidence in support of the usefulness of trait activation as a predictor of trait-behavior relationships (Tett & Guterman, 2000). Moreover, several studies have suggested dimension-based performance links to personality variables. Though results have spawned a few discouraging result (Chan, 1996; Fleenor, 1997), some have been quite successful in establishing expected relationships between assessor ratings within a nomological network of related variables. For instance, in a study conducted by Shore, Thornton, & Shore (1990), 441 candidates were assessed on 11 dimensions (grouped as either interpersonal-style or performance-style) in three leaderless group discussions and an interview, and scores were collected on

several cognitive ability tests and the 16-personality factor. Results were especially noteworthy for interpersonal style dimensions. Specifically, two out of three interpersonal-style dimensions (e.g., amount of participation and impact) were found to relate more strongly to conceptually similar than dissimilar 16 personality factor scales. Moreover, within the groupings, dimension scores were on average more highly correlated than across the two groups (performance style mean r = .59, interpersonal style mean r = .51, and across group mean r = .46).

In a related study, Thornton, Tziner, Dahan, Clevenger, & Meir (1997) correlated final dimension/attribute ratings from 382 mid-level managers in a manufacturing organization with results from tests and inventories measuring sixteen comparable attributes. These tests and inventories included the 16-personality factor, the Minnesota Multiphasic Personality Inventory, two projective tests (the Rosenzweig Picture Interpretation Test and the Miner Sentence Completion Test), the Bender Gestalt, and the Reddin Managerial Style Questionnaire. They also found mixed results. Confirmatory factor analysis following the proposed structure revealed poor fit of the data (GFI = .71, RMSR = .16). However, when comparing assessor ratings with psychologists' evaluations, assessor ratings correlated more highly with comparable sets of measures than noncomparable sets for five of seven comparisons ($p < .05$). As an example, with regard to the dimension "Standards of High Performance," assessor ratings correlated more strongly with psychologists' judgments of perseverance and dedication (r = .42) than with judgments of creativity, maturity and stability, independence, and energy level (average r = .06).

All in all, there appears to be a potential link between the traits and dimensions that are activated within given situations or exercises. An alternative possibility is that the activation of a personality construct that is salient within a given exercise may manifest itself in behavior relevant to certain dimensions of performance, but to a lesser effect in other dimensions, effecting differential dimension activation. The questions consequently arises, "Can we obtain stronger construct- and criterion-related validity when designing exercises to elicit personality traits relevant to performance dimension or are we better off designing exercises based on dimension themselves as done traditionally?" and, "Can important personality characteristics necessary for job performance be identified more reliably than dimensions of performance (the typical procedure used when analyzing jobs to create exercises)?"

In Conclusion

For decades, researchers have utilized the multi-trait multi-method framework for assessing the construct-related validity of assessment center ratings and have been discouraged by results. However, a careful examination of the exercise design strategy suggests that the multi-trait multi-method analysis is inappropriate for such purposes. This study was the first attempt to assess assessment center construct-related validity in the context of exercise design strategy. The obtained results provided initial support for the Dimension Activation Theory as an explanatory tool for understanding past findings and in doing so, shed some light into the black box of the person-situation (or dimension-exercise) interaction in assessment centers. Perhaps most importantly, this study highlights the need to develop a better understanding of exercise design strategies and their impact on assessment center construct- and criterion-related validity.

REFERENCES

Archambeau, D. J. (1979). Relationships among skill ratings assigned in an assessment center. Journal of Assessment Center Technology, 2, 7-20.

Arthur, W. A., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. Journal of Management, 26, 813-835.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Bulletin, 81, 506-520.

Bem, D. J., & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. Psychological Review, 85, 485-501.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74, 478-494.

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. Journal of Applied Psychology, 74, 957-963.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. Organizational Research Methods, 2(1), 49-68.

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. Journal of Applied Psychology, 72, 463-474.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Chan, D. (1996). Criterion and construct validation of an assessment center. Journal of Occupational and Organizational Psychology, 69, 167-181.

Cohen, J. (1969). Statistical Power Analyses for the Behavioral Sciences. New York: Academic Press, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. Journal of Social Behavior and Personality, 12, 85-108.

Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. Psychological Bulletin, 98, 513-537.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. Journal of Business and Psychology, 10, 319-333.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, 74, 611-618.

Gaugler, B. B., Rosenthal, G. C., Thornton, G. C., Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72, 493-511.

Guion, R. M. (1987). Changing views for personnel selection research. Personnel Psychology, 40, 199-213.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. Personnel Psychology, 55, 137-163.

Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? Journal of Applied Psychology, 78(4), 675-678.

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. Journal of Applied Social Psychology, 23, 140-155.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. In Riggio, R. E. & Mayes, B. T. (Eds.), Assessment Centers: Research and Applications [Special Issue]. Journal of Social Behavior and Personality, 12, 13-52.

International Task Force on Assessment Center Guidelines. (2000). Guidelines and ethical consideration for assessment center operations. Public Personnel Management, 29(3), 315-331.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69, 85-98.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behavioral Research, 35(1), 1-19.

Johnson, J. W., & LeBreton, J. M. (2002). History and Use of Relative Importance Indices in Organizational Research. Paper submitted to Organizational Research Methods.

Jones, R. G. (1992). Construct validation of assessment center final dimension ratings: Definition and measurement issues. Human Resource Management Review, 2, 195-220.

Joyce, L., Thayer, P., & Pond, S. (1994). Managerial functions: An alternative to traditional assessment center dimensions? Personnel Psychology, 47, 109-121.

Kanji, G. K. (1993). Statistical Tests. London: Sage.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. American Psychologist, 43, 23-34.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. Journal of Applied Psychology, 78, 988-993.

Kleinmann, M., & Kohller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. Journal of Social Behavior and Personality, 12, 65-84.

Kleinmann, M., Kuptsch, C., & Koller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centers. Applied Psychology: An International Review, 45, 67-84.

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. Personnel Psychology, 40, 243-257.

Klimoski, R. J. & Strickland, W. J. (1977). Assessment centers – valid or merely prescient. Personnel Psychology, 30, 353-361.

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. Journal of Social Behavior and Personality, 12, 129-144.

Kuptsch, C., Kleinmann, M., & Koller, O. (1998). The chameleon effect in assessment centers: The influence of cross-situational behavioral consistency on the convergent validity of assessment centers. Journal of Social Behavior and Personality, 13(1), 103-116.

Ladd, R. T., Atchley, E. K., & Burgess, J. R. D. (2001). What good is importance if you don't know how to use it? A comparison of various relative importance indices and a heuristic for their use in selecting predictor variables. Paper presented at the Annual Conference of the Society for Industrial/Organizational Psychology.

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoreson, J. D. (in press). Assessor Cognitive Processes in an Operational Assessment Center.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. Journal of Applied Psychology, 79(3), 332-340.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. S., French, N. R., & Smith, D. E. (2000). Assessment center exercises represent cross-situational specificity, not method bias. Human Performance, 13, 323-353.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. International Journal of Selection and Assessment, 6, 141-152.

Lievens, F. (2001). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. Journal of Organizational Behavior, 22, 203-221.

Lievens, F. & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. Journal of Applied Psychology, 86(6), 1202-1222.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), International Review of Industrial and Organizational Psychology, Vol. 16, pp. 245-286.

Lord, C. G. (1982). Predicting behavioral consistency from an individual's perception of situational similarities. Journal of Personality and Social Psychology, 44, 1076-1088.

Magnusson, D. (1982). Toward a psychology of situations: An interactional perspective. Hillsdale, NJ: Erlbaum.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (ED.), Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: Sage Publications.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross situational consistency. Psychological Review, 89, 730-755

Mischel W, & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, ad invariance in personality structure, Psychological Review, 102, 246-268.

Mischel, W. & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. Annual Review of Psychology, 49, 229-258.

Neidig, R. D., Martin, J. C., & Yates, R. E. (1979). The contribution of exercise skill ratings to final assessment center evaluations. Journal of Assessment Center Technology, 2, 21-23.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness, Journal of Applied Psychology, 69, 182-186.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models, Fourth Edition. McGraw-Hill, Boston.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. Personnel Psychology, 43, 71-84.

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centers: Dimensions into exercises won't go. Journal of Occupational Psychology, 60, 187-195.

Russell, C. (1985). Individual decision processes in an assessment center. Journal of Applied Psychology, 70, 737-746.

Russell, C. (1987). Person characteristics versus role congruency explanations for assessment center ratings. Academy of Management Journal, 4, 817-826.

Russell, C. J., & Dom, D. R. (1995). Two field test of an explanation of assessment center validity. Journal of Occupational and Organizational Psychology, 68, 25-47.

Sackett, P. R. (1982). A critical look at some common beliefs about assessment centers. Public Personnel Management, 11, 140-147.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401-410.

Sackett, P. R., & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. Organizational Behavior and Human Performance, 23, 120-137.

Sackett P. R., & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. Journal of Business and Psychology, 3, 214-229.

Sackett, P. R., & Tuzinski, K. A. (2002). The role of dimensions and exercises in assessment center judgments. In London M (Ed.), How people evaluate others in organizations (pp. 111-129). Mahwah, NJ:LEA.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. Journal of Applied Psychology, 62, 171-176.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 32-41.

Schmitt, N., Schneider, J. R., & Cohen, S. A. (1990). Factors affecting validity of a regionally administered assessment center. Personnel Psychology, 43, 1-12.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimensions and exercise constructs. Journal of Applied Psychology, 77, 32-41.

Shoda Y, Mischel, W., & Wright, J. C. (1989). Intuitive interactionism in person perception: Effects of situation-behavior relations on dispositional judgments. Journal of Personality and Social Psychology, 56, 41-52.

Shoda, Y., Mischel, W., & Wright J. C. (1993). The role of situational demands and cognitive competencies in behavior organization and personality coherence. Journal of Personality and Social Psychology, 65, 1023-1035.

Shore, T., Thronton, G., & Shore, L. (1990). Construct validity of two categories of assessment center dimension ratings. Personnel Psychology, 39, 565-578.

Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L., Jr. (1986). Influence of assessment center methods on assessors' ratings. Personnel Psychology, 39, 565-578.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. A. (1997). A survey of assessment center practices in organizations in the United States. Personnel Psychology, 50, 71-90.

Sulsky, L. M., & Balzer, W. K. (1988). The meaning and measurement of performance rating accuracy: Some methodological concerns. Journal of Applied Psychology, 73, 497-506.

Task Force on Assessment Center Guidelines. (1989). Guidelines and ethical considerations for assessment center operations. Public Personnel Management, 18, 457-470.

Tett, R. P., & Burnett, D. B. (in press). A personality trait-based interactionist model of job performance. Journal of Applied Psychology.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross situational consistency: Testing a principle of trait-activation. Journal of Research in Personality, 34, 397-423.

Tett, R. P., & Schleicher, D. J. (19??). Assessment center dimensions as "traits": New concepts in assessment center design.

Thornton, G. C. III, & Byham, W. C. (1982). Assessment centers and managerial performance. New York: Academic Press.

Thornton, G. C., III, Tziner, A., Dahan, M., Clevenger, J. P., & Meir, E. (1997). Construct validity of assessment center judgments: Analysis of the behavioral reporting method. Journal of Social Behavior and Personality, 12(5), 109-128.

Torgerson, W. S. (1958). Theory and Methods of Scaling. London: John Wiley & Sons, Inc.

Turnage J. J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. Organizational Behavior and Human Performance, 30, 174-200.

Turnage, J. J, & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. Journal of Applied Psychology, 69(4), 595-602.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company.

Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. Journal of Organizational Behavior, 13, 357-367.

Woehr, D. J., & Arthur, W. Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. Journal of Management, 29(2), 231-258.

Zedeck, S. (1986). A process analysis of the assessment center method. Research in Organizational Behavior, 8, 259-296.

APPENDICES

A. Assessment Center Description

Description of Performance Dimensions

Oral Communication
Oral communication is effective expression in one-on-one or group situations. It includes delivery, clarity of ideas, and speaking with enthusiasm and confidence.

Analysis
Analysis refers to the ability to identify problems, secure relevant information, relate data from different sources, and identify causes of problems.

Judgment
Judgment refers to the ability to develop alternative courses of action and make decision based on logical assumptions that reflect factual information. Judgment also includes providing rational for decisions and recommendations.

Planning and Organizing
Planning and organizing refers to the ability to establish a course of action to accomplish a specific goal. It includes factors such as setting priorities and making appropriate allocation of time and resources.

Decisiveness
Decisiveness is the willingness to make decisions, render judgments, take action, or commit one's self. It also includes firmly stating one's opinion on an issue.

Delegation
Delegation refers to utilizing subordinates effectively. It implies direction, accountability, and control. Good delegation is clear, best suited to the individual, includes deadlines, and sets limits on authority.

Initiative
Initiative is the extent to which an individual is a self-starter and actively attempts to influence events to achieve goals.

Leadership
Leadership refers to utilizing appropriate interpersonal styles and methods in guiding individuals (subordinates/peers/superiors) or groups toward task accomplishment.

Coaching
Coaching is defined as the extent to which the individual offers clear, concrete direction and guidance for dealing with the problem situation.

Team Building
Team Building refers to the ability to work effectively as part of a group, and the willingness to act as part of a team and accept team-based decisions.

Confrontation
Confrontation is defined as the ability and willingness to disagree or express opposing viewpoints in a tactful style. It also includes the willingness to assert and defend one's position even when challenged.

Sensitivity
Sensitivity is demonstrated with actions that indicate a consideration of the feelings and needs of others. A highly sensitive individual is not brash, rude, or threatening, asks for the opinions of others, and gives encouragement.

Customer Orientation
Customer Orientation is defined as the extent to which a person places customer service and customer satisfaction as a high priority, and acts to serve customers in a way which will yield satisfaction.

Assessment Center Exercises

In-Basket Exercise

This exercise is a partial simulation of administrative tasks associated with upper-level managerial jobs. It requires quick analysis and action on a number of complex, high priority items. The task requires that the applicant shows initiative and begins implementing a strategic plan. Assessing priorities, making decisions on limited information, recognizing conflicting information, and managing a very busy schedule are key elements to this exercise.

Simulation Role Plays

This exercise simulates two types of interpersonal situations that might be expected in a managerial position. Two role players, each with a defined script, serve as subordinates of the applicant. Successful resolution of each task depends on being able to correctly assess the nature of the problem from both the materials provided and information provided by the subordinate, assessing the problem in an appropriate manner, and effectively counseling the subordinate as necessary. Good interpersonal skills are required to perform well in this exercise.

Group Decision Making Exercise

This task requires the applicant to participate on a school board committee allocating a substantial sum of money to various needs of a school system. This exercise contains two major parts. First, the applicant is required to individually assess the situation and decide how the funds should be allocated among a number of proposals. Next, the "school board committee," which consists of three role players and the applicant, must arrive at a consensus decision on how funds should be distributed. This problem was designed to be relatively independent of the particular expertise of any applicant. Successful completion of this exercise requires that the applicant be able to communicate in a group setting, orally support his or her ideas, and be able to exert influence on others.

B. Dimensions Assessed in Each Exercise

Dimensions Assessed in Each Exercise

	OC	*WC*	*A*	*J*	*PO*	*Dec*	*Del*	*L*	*In*	*Coa*	*TB*	*Con*	*Sen*	*CO*
IB			X	X	X	X	X		X		X			X
Group	X		X	X				X	X		X	X	X	
RP1	X		X	X		X	X	X		X		X	X	
RP2	X		X	X		X		X	X	X		X	X	

Note: IB is the In-Basket exercise, Group is the Group Decision Making Task, RP1 and RP2 are Role Playing exercises, OC is oral communication, A is anlaysis, J is judgment, PO is planning and organizing, Dec is decisiveness, Del is delegation, L is leadership, In is initiative, Coa is coaching, TB is team building, Con is confrontation, S is sensitivity, and CO is customer orientation.

C. Dimension Activation Inventory

Instructions: Please consider each paring of the dimensions typically rated in each exercise. Distribute 100 points among each pairing with the dimension having the level of ***highest*** dimension activation potential (see below for definition) among the two choices receiving the highest number of points. Repeat for each pairing.

Dimension Activation Potential: The applicability of a particular dimension to the exercise at hand. That is, the importance of performance in that dimension to overall exercise performance.

CASE ANALYSIS EXERCISE

1.	_____ Analysis	_____ Judgment
2.	_____ Written Communication	_____ Analysis
3.	_____ Planning and Organizing	_____ Decisiveness
4.	_____Judgment	_____ Written Communication
5.	_____ Planning and Organizing	_____ Written Communication
6.	_____ Analysis	_____ Decisiveness
7.	_____ Judgment	_____ Planning and Organizing
8.	_____ Written Communication	_____ Decisiveness
9.	_____ Planning and Organizing	_____ Analysis
10.	_____ Decisiveness	_____ Judgment

IN-BASKET EXERCISE

1. _____ Delegation	_____ Analysis
2. _____ Team Building	_____ Sensitivity
3. _____ Analysis	_____ Team Building
4. _____ Written Communication	_____ Analysis
5. _____ Decisiveness	_____ Planning and Organizing
6. _____ Sensitivity	_____ Written Communication
7. _____ Judgment	_____ Delegation
8. _____ Sensitivity	_____ Planning and Organization
9. _____ Written Communication	_____ Decisiveness
10. _____ Delegation	_____ Team Building
11. _____ Planning and Organizing	_____ Written Communication
12. _____ Analysis	_____ Decisiveness
13. _____ Team Building	_____ Judgment
14. _____ Sensitivity	_____ Delegation
15. _____ Delegation	_____ Planning and Organizing
16. _____ Team Building	_____ Decisiveness
17. _____ Sensitivity	_____ Judgment
18. _____ Decisiveness	_____ Sensitivity
19. _____ Judgment	_____ Analysis
20. _____Team Building	_____ Written Communication
21. _____ Planning and Organizing	_____ Judgment
22. _____ Judgment	_____ Written Communication

23. _____ Analysis	_____ Planning and Organizing
24. _____ Judgment	_____ Decisiveness
25. _____ Decisiveness	_____ Delegation
26. _____ Planning and Organizing	_____ Team Building
27. _____ Analysis	_____ Sensitivity
28. _____ Written Communication	_____ Delegation
29. _____ Written Communication	_____ Initiative
30. _____ Analysis	_____ Customer Orientation
31. _____ Analysis	_____ Initiative
32. _____ Customer Orientation	_____ Team Building
33. _____ Planning and Organizing	_____ Initiative
34. _____ Initiative	_____ Customer Orientation
35. _____ Customer Orientation	_____ Decisiveness
36. _____ Delegation	_____ Initiative
37. _____ Customer Orientation	_____ Written Communication
38. _____ Judgment	_____ Initiative
39. _____ Customer Orientation	_____ Delegation
40. _____ Initiative	_____ Decisiveness
41. _____ Initiative	_____ Team Building
42. _____ Customer Orientation	_____ Judgment
43. _____ Planning and Organizing	_____ Customer Orientation

SIMULATION ROLE PLAY (strong role/CH)

1. _____ Leadership	_____ Delegation
2. _____ Analysis	_____ Coaching
3. _____ Delegation	_____ Analysis
4. _____ Coaching	_____ Judgment
5. _____ Oral Communication	_____ Decisiveness
6. _____ Leadership	_____ Coaching
7. _____ Decisiveness	_____ Judgment
8. _____ Confrontation	_____ Sensitivity
9. _____ Coaching	_____ Decisiveness
10. _____ Oral Communication	_____ Leadership
11. _____ Decisiveness	_____ Analysis
12. _____ Confrontation	_____ Delegation
13. _____ Judgment	_____ Oral Communication
14. _____ Delegation	_____ Coaching
15. _____ Analysis	_____ Sensitivity
16. _____ Oral Communication	_____ Coaching
17. _____ Judgment	_____ Confrontation
18. _____ Oral Communication	_____ Delegation
19. _____ Leadership	_____ Analysis
20. _____ Confrontation	_____ Analysis
21. _____ Delegation	_____ Sensitivity
22. _____ Oral Communication	_____ Sensitivity

23. _____ Judgment	_____ Leadership
24. _____ Analysis	_____ Oral Communication
25. _____ Confrontation	_____ Coaching
26. _____ Sensitivity	_____ Decisiveness
27. _____ Analysis	_____ Judgment
28. _____ Leadership	_____ Decisiveness
29. _____Confrontation	_____ Decisiveness
30. _____ Sensitivity	_____ Judgment
31. _____ Leadership	_____ Confrontation
32. _____Decisiveness	_____ Delegation
33. _____ Leadership	_____ Sensitivity
34. _____ Judgment	_____ Delegation
35. _____ Confrontation	_____ Oral Communication
36. _____ Sensitivity	_____ Coaching

SIMULATION ROLE PLAY (weak role/BM)

1. _____ Sensitivity
2. _____Oral Communication
3. _____ Coaching
4. _____Confrontation
5. _____ Leadership
6. _____ Decisiveness
7. _____ Sensitivity
8. _____ Leadership
9. _____ Delegation
10. _____ Sensitivity
11. _____ Coaching
12. _____ Analysis
13. _____ Delegation
14. _____ Coaching
15. _____ Coaching
16. _____ Decisiveness
17. _____ Delegation
18. _____ Oral Communication
19. _____ Judgment
20. _____ Confrontation
21. _____ Decisiveness
22. _____ Decisiveness

_____ Judgment
_____ Delegation
_____ Sensitivity
_____ Oral Communication
_____ Analysis
_____ Judgment
_____ Delegation
_____ Confrontation
_____ Confrontation
_____ Oral Communication
_____ Oral Communication
_____ Decisiveness
_____ Judgment
_____ Delegation
_____ Confrontation
_____ Leadership
_____ Analysis
_____ Leadership
_____ Coaching
_____ Sensitivity
_____ Coaching
_____ Delegation

23. _____ Analysis	_____ Judgment
24. _____ Leadership	_____ Delegation
25. _____ Confrontation	_____ Analysis
26. _____ Decisiveness	_____ Oral Communication
27. _____ Leadership	_____ Coaching
28. _____ Sensitivity	_____ Decisiveness
29. _____ Analysis	_____ Oral Communication
30. _____ Judgment	_____ Leadership
31. _____ Decisiveness	_____ Confrontation
32. _____ Analysis	_____ Sensitivity
33. _____ Confrontation	_____ Judgment
34. _____ Leadership	_____ Sensitivity
35. _____ Oral Communication	_____ Judgment
36. _____ Analysis	_____ Coaching
37. _____ Oral Communication	_____ Initiative
38. _____Analysis	_____ Initiative
39. _____ Judgment	_____ Initiative
40. _____Decisiveness	_____ Initiative
41. _____ Leadership	_____ Initiative
42. _____ Coaching	_____ Initiative
43. _____ Confrontation	_____ Initiative
44. _____ Sensitivity	_____ Initiative

GROUP DECISION TASK

1. _____ Leadership	_____ Analysis
2. _____ Team Building	_____ Judgment
3. _____ Judgment	_____ Oral Communication
4. _____ Oral Communication	_____ Sensitivity
5. _____ Team Building	_____ Analysis
6. _____ Confrontation	_____ Sensitivity
7. _____ Oral Communication	_____ Team Building
8. _____ Analysis	_____ Judgment
9. _____ Oral Communication	_____ Analysis
10. _____ Judgment	_____ Leadership
11. _____ Sensitivity	_____ Analysis
12. _____ Leadership	_____ Team Building
13. _____ Confrontation	_____ Judgment
14. _____ Analysis	_____ Confrontation
15. _____ Team Building	_____ Sensitivity
16. _____ Oral Communication	_____ Leadership
17. _____ Judgment	_____ Sensitivity
18. _____ Leadership	_____ Sensitivity
19. _____ Leadership	_____ Confrontation
20. _____ Analysis	_____ Oral Communication
21. _____ Confrontation	_____ Team Building
22. _____ Oral Communication	_____ Initiative

23. _____ Analysis _____ Initiative

24. _____ Judgment _____ Initiative

25. _____ Leadership _____ Initiative

26. _____ Team Building _____ Initiative

27. _____ Confrontation _____ Initiative

28. _____ Sensitivity _____ Initiative

DEMOGRAPHIC INFORMATION

1. AGE _____

2. RACE White African American Hispanic Asian Other _____

3. GENDER Male Female

4. LEVEL OF I/O GRADUATE WORK COMPLETED _____ years

5. YEARS INVOLVED IN EMPLOYEE ASSESSMENT _____ years

6. NUMBER OF ASSESSMENT CENTERS RATED _____

7. NUMBER OF CANDIDATES RATED _____

8. YEARS SUPERVISORY EXPERIENCE _____ years

D. Variance Estimates by Dimension

Variance Estimates by Dimension

Exercise	*Oral Communication*	*Analysis*	*Judgment*	*Planning Organizing*	*Decisiveness*	*Delegation*	*Leadership*	*Initiative*	*Coaching*	*Team Building*	*Confrontation*	*Sensitivity*	*Customer Orientation*
In-Basket		**.33**	**.37**	.18	**.53**	.24		.31		.14			.31
Role Play 1	.04	**.17**	**.28**		**.22**	.18	**.23**		.14		**.14**	.15	
Role Play 2	.07	**.37**	**.27**		.20		**.42**	.34	**.34**		.17	.21	
Group	.05	**.10**	**.13**				**.26**	.16		**.12**	.10	.08	

Note: All activated dimensions are in boldface.

VITA

Michelle A. Bush was born on February 28, 1976 in Des Moines, Iowa. A move at an early age relocated her family to nearby Cedar Rapids, Iowa, where she was raised by her parents, Alan and Dee, along with her sister Pam. After graduating from Linn-Mar High School in 1994, Michelle attended the University of Northern Iowa as a Presidential Scholar. She graduated magna cum laude with a Bachelor of Arts in Psychology in May, 1998. Immediately following graduation, she moved to Knoxville, Tennessee to pursue a doctoral degree in Industrial/Organizational Psychology from the University of Tennessee. This degree was presented in December of 2003.

While studying at the University of Tennessee, Michelle earned numerous opportunities to apply her studies in a variety of settings. In particular, she served as Coordinator for the Tennessee Assessment Center between the years of 2000 to 2001, and was involved as an assessor/role player between the years of 1999 and 2003. In addition, Michelle was involved with executive and managerial leadership development for the University of Tennessee Physician/Executive Masters of Business Administration programs. Moreover, she has consulted with various businesses include Ruby Tuesdays, Inc. and the Knoxville Chamber of Commerce, and currently serves as the Developmental Co-Coordinator at Covenant Health Systems Thompson Cancer Survival Center.

Michelle has also had the opportunity to present research studies at the annual Society for Industrial and Organizational Psychology Convention as well as the Industrial/Organizational and Organizational Behavior Convention. Her research was

additionally recognized and showcased at an academic research consortium within the University of Tennessee College of Business.